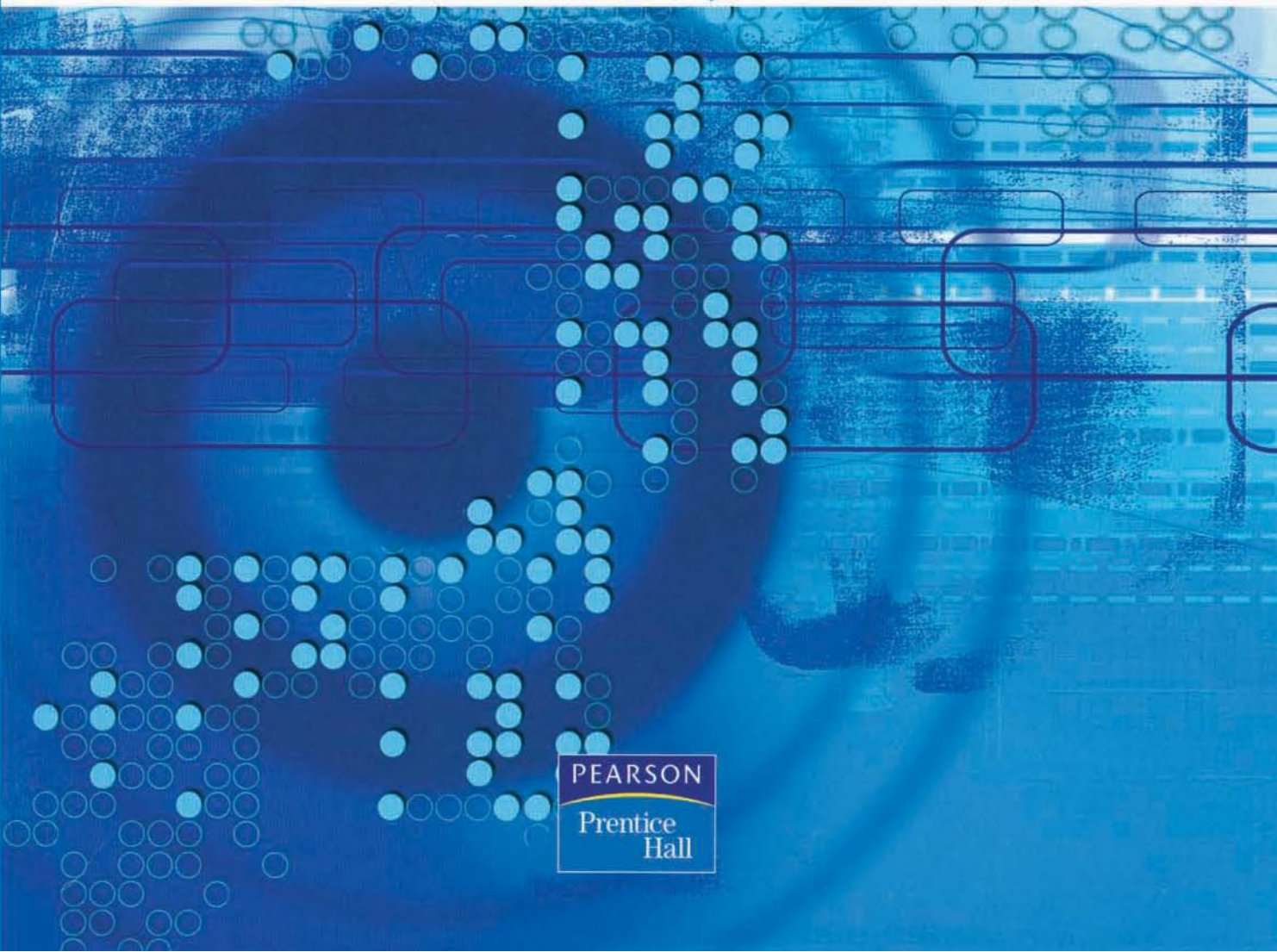
PRENTICE PRÁCTICA

Incluye CD-ROM

Muestreo estadístico

Conceptos y problemas resueltos

César Pérez López



PEARSON
Prentice Hall

Muestreo estadístico

Conceptos y problemas resueltos

Muestreo estadístico

Conceptos y problemas resueltos

CÉSAR PÉREZ LÓPEZ
Universidad Complutense de Madrid

Instituto de Estudios Fiscales

PEARSON
Prentice Hall

Madrid • México • Santafé de Bogotá • Buenos Aires • Caracas • Lima • Montevideo
San Juan • San José • Santiago • São Paulo • White Plains

Datos de catalogación bibliográfica

MUESTREO ESTADÍSTICO
CONCEPTOS Y PROBLEMAS RESUELTOS
CÉSAR PÉREZ LÓPEZ

PEARSON EDUCACIÓN, S.A., Madrid, 2005

ISBN: 84-205-4411-6
Materia: Estadística Matemática, 519.2

Formato: 195 × 270 Páginas: 392

Todos los derechos reservados.
Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con autorización de los titulares de propiedad intelectual. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual *(arts. 270 y sgts. Código Penal).*

DERECHOS RESERVADOS
© 2005 por PEARSON EDUCACIÓN, S.A.
Ribera del Loira, 28
28042 Madrid (España)

MUESTREO ESTADÍSTICO
CONCEPTOS Y PROBLEMAS RESUELTOS
CÉSAR PÉREZ LÓPEZ

ISBN: 84-205-4411-6
Depósito Legal: M. 18.283-2005

PEARSON PRENTICE HALL es un sello editorial autorizado de PEARSON EDUCACIÓN, S.A.

Equipo editorial:
Editor: Miguel Martín-Romo
Técnico editorial: Marta Caicoya

Equipo de producción:
Director: José Antonio Clares
Técnico: José Antonio Hernán

Diseño de cubierta: Equipo de diseño de Pearson Educación, S.A.

IMPRESO EN MÉXICO - PRINTED IN MEXICO

A mis niñas

CONTENIDO

INTRODUCCIÓN xi

CAPÍTULO 1. MUESTREO ESTADÍSTICO: CONCEPTOS, ESTIMADORES Y SU DISTRIBUCIÓN 1

Conceptos iniciales en la teoría del muestreo 3
Muestreo y estimadores. Distribuciones en el muestreo 4
Propiedades y precisión de los estimadores. Comparación de estimadores 6
Estimación por intervalos de confianza 8
Problemas resueltos 10
Ejercicios propuestos 47

CAPÍTULO 2. MÉTODOS GENERALES DE SELECCIÓN DE MUESTRAS. ESTIMACIÓN Y ERRORES 49

Selección con y sin reposición. Probabilidades iguales y desiguales 51
Estimación puntual y formación general de estimadores 51
Muestreo con reposición y probabilidades desiguales. Estimador de Hansen Hurwitz 54
Muestreo con reposición y probabilidades proporcionales a los tamaños. Métodos especiales de selección 55
Muestreo sin reposición y probabilidades desiguales. Estimador de Horvitz Thompson 56
Muestreo sin reposición y probabilidades proporcionales a los tamaños. Métodos especiales de selección 57
Método de Montecarlo 62
Problemas resueltos 64
Ejercicios propuestos 108

CAPÍTULO 3. MUESTREO ALEATORIO SIMPLE SIN Y CON REPOSICIÓN. SUBPOBLACIONES 109

Muestreo aleatorio simple sin reposición. Especificaciones 111
Estimadores, varianzas y estimación de varianzas 112
Tamaño de la muestra 114

Muestreo aleatorio simple con reposición. Estimadores 118
Varianzas y su estimación con reposición 119
Tamaño de la muestra con reposición 120
Comparación entre muestreo aleatorio sin y con reposición 121
Subpoblaciones 122
Problemas resueltos 124
Ejercicios propuestos 145

CAPÍTULO 4. MUESTREO ESTRATIFICADO SIN Y CON REPOSICIÓN 147

Concepto de muestreo estratificado 149
Muestreo estratificado sin reposición. Estimadores y errores 150
Muestreo estratificado con reposición. Estimadores y errores. 151
Afijación de la muestra. Tipos de afijación y errores de los estimadores para muestreo sin reposición 152
Afijación de la muestra. Tipos de afijación y errores de los estimadores para muestreo con reposición 155
Tamaño de la muestra para muestreo sin reposición 156
Tamaño de la muestra para muestreo con reposición 156
Comparación de eficiencias en muestreo estratificado 157
Postestratificación 159
Problemas resueltos 161
Ejercicios propuestos 195

CAPÍTULO 5. MUESTREO SISTEMÁTICO 197

Muestreo sistemático. Especificaciones 199
Estimadores y varianzas 200
Relación entre el muestreo sistemático y el muestreo aleatorio simple 203
Relación entre el muestreo sistemático y el muestreo estratificado 203
Estimación de varianzas 205
Relación entre el muestreo sistemático y el muestreo por conglomerados 206
Problemas resueltos 207
Ejercicios propuestos 224

CAPÍTULO 6. MUESTREO POR MÉTODOS INDIRECTOS. RAZÓN, REGRESIÓN Y DIFERENCIA 225

Estimadores no lineales 227
Muestreo por métodos indirectos. El estimador de razón 229
Estimaciones de los parámetros poblacionales basadas en la razón y errores 233
Estimaciones por regresión y errores 234
Estimaciones por diferencia y errores 237
Estimadores de razón en el muestreo estratificado 238
Estimadores de regresión en el muestreo estratificado 245
Problemas resueltos 250
Ejercicios propuestos 271

CAPÍTULO 7. MUESTREO UNIETÁPICO DE CONGLOMERADOS 273

Muestreo unietápico de conglomerados. Estimadores para conglomerados del mismo tamaño y probabilidades iguales 275
Varianza de los estimadores. Coeficiente de correlación intraconglomerados. Estimación de varianzas 276
Muestreo de conglomerados del mismo tamaño con reposición. Varianzas de los estimadores y estimación de las varianzas 280
Muestreo unietápico de conglomerados de distinto tamaño 281
Muestreo unietápico de conglomerados de distinto tamaño con probabilidades desiguales 283
Tamaño de la muestra 285
Problemas resueltos 286
Ejercicios propuestos 297

CAPÍTULO 8. MUESTREO BIETÁPCO DE CONGLOMERADOS 299

Muestreo bietápico de conglomerados. Estimadores para probabilidades iguales y conglomerados del mismo tamaño 301
Varianzas y su estimación en muestreo bietápico con probabilidades iguales y conglomerados del mismo tamaño 301
Muestreo bietápico de conglomerados de distinto tamaño y probabilidades iguales 304
Tamaño de la muestra en muestreo bietápico 307
Muestreo bietápico con probabilidades desiguales y con reposición en 1ª etapa. Estimadores, varianzas y su estimación 308
Muestreo bietápico con probabilidades desiguales y sin reposición en 1ª etapa. Estimadores, varianzas y su estimación 310
Muestreo polietápico 312
Diseños complejos. Muestreo bietápico con estratificación en primera etapa 313
Problemas resueltos 314
Ejercicios propuestos 327

CAPÍTULO 9. MUESTREO BIFÁSICO Y MUESTREO EN OCASIONES SUCESIVAS 329

Muestreo bifásico 331
Muestreo bifásico para estratificación. Estimadores, varianzas y estimación de varianzas 332
Muestreo bifásico para estimadores de razón 336
Muestreo bifásico para estimadores de regresión 337
Muestreo bifásico para estimadores de diferencia 338
Mestreo en ocasiones sucesivas 338
Estimadores de mínima varianza en el muestreo en ocasiones sucesivas 341
Problemas resueltos 344
Ejercicios propuestos 350

CAPÍTULO 10. MUESTREO ESTADÍSTICO MEDIANTE SPSS 351

SPSS y el muestreo estadístico 353
Diseños complejos y el asistente de muestreo. Creación de un nuevo plan de muestreo 354
Asistente de muestreo: Modificar un plan existente 362
Asistente de muestreo: Ejecutar un plan de muestreo dado 364
Preparación de una muestra compleja para su análisis: Creación de un nuevo plan de análisis 364
Preparación de una muestra compleja para su análisis: Modificar un plan de análisis existente 368
Cálculos en muestras complejas: frecuencias, descriptivos, tablas de contingencia y razones 368

INTRODUCCIÓN

La finalidad esencial de este libro es presentar las técnicas de muestreo estadístico en su faceta práctica. Cada capítulo comienza con una breve exposición de los conceptos teóricos a utilizar en los problemas con el objetivo de que no sea necesario recurrir a textos externos para comprender las herramientas utilizadas en la solución de los ejercicios. Además, determinados ejercicios se refuerzan con aplicaciones informáticas para obtener la solución. En particular se utilizan Excel y SPSS.

Los más de 150 problemas que contiene el texto, así como los conceptos teóricos, se dirigen tanto a docentes como a estudiantes universitarios de todos los niveles que imparten o cursan la materia de muestreo estadístico. El libro es también de utilidad para los profesionales de la economía, biología, botánica, zoología, marketing, auditoría, agronomía, comercio, transporte, medicina, control de calidad, etc. En general puede utilizarse en todos los sectores en los que se aplican las técnicas de muestreo.

En cuanto al contenido, se comienza exponiendo los conceptos iniciales de la teoría del muestreo, para facilitar la situación del lector en el contexto de la teoría de muestras moderna. A continuación se presentan los métodos básicos para seleccionar la muestra y se desarrollan los diferentes tipos de muestreo, como muestreo aleatorio simple, muestreo estratificado, muestreo sistemático, métodos indirectos de estimación por razón, regresión y diferencia, muestreo por conglomerados unietápico, bietápico y polietápico, los procedimientos para el muestreo bifásico y los problemas peculiares de las encuestas repetidas.

Los problemas suelen adecuarse en lo posible a situaciones prácticas y la metodología pretende mantener la secuencia conceptos → aplicaciones, muy útil en los métodos de enseñanza modernos. Comenzar presentando los temas de forma teórica, para a continuación resolver ejercicios prácticos que ilustran los métodos teóricos, cuya resolución suele apoyarse en la medida de lo posible en las herramientas informáticas más adecuadas, es la secuencia más lógica a seguir en la didáctica de esta materia.

CAPÍTULO

1

MUESTREO ESTADÍSTICO: CONCEPTOS, ESTIMADORES Y SU DISTRIBUCIÓN

OBJETIVOS

1. Presentar el concepto de muestreo estadístico en poblaciones finitas.
2. Distinguir claramente los conceptos de población, marco y muestra.
3. Introducir el concepto de estimador y su distribución en el muestreo.
4. Analizar las propiedades de los estimadores.
5. Estudiar la precisión de los estimadores.
6. Comparar estimadores.
7. Cuantificar la precisión de los estimadores.
8. Comprender el concepto de estimación mediante intervalos de confianza.
9. Analizar la influencia del sesgo en la estimación por intervalos de confianza.
10. Analizar la influencia de la normalidad en la estimación por intervalos de confianza.
11. Realizar la estimación mediante intervalos de confianza.

ÍNDICE

1. Conceptos iniciales en la teoría del muestreo.

2. Muestreo y estimadores. Distribuciones en el muestreo.

3. Propiedades y precisión de los estimadores. Comparación de estimadores.

4. Estimación por intervalos de confianza.

5. Problemas resueltos.

6. Ejercicios propuestos.

CONCEPTOS INICIALES EN LA TEORÍA DEL MUESTREO

Al hablar de *métodos de muestreo* nos referimos al conjunto de técnicas estadísticas que estudian la forma de seleccionar una *muestra lo suficientemente representativa* de una población cuya información permita inferir las propiedades o características de toda la población cometiendo un *error medible y acotable*. A partir de la muestra, seleccionada mediante un determinado método de muestreo, se estiman las características poblacionales (media, total, proporción, etc.) con un error cuantificable y controlable. Las estimaciones se realizan a través de funciones matemáticas de la muestra denominadas *estimadores*, que se convierten en variables aleatorias al considerar la variabilidad de las muestras. Los errores se cuantifican mediante varianzas, desviaciones típicas o errores cuadráticos medios de los estimadores, que miden la precisión de éstos. La metodología que permite inferir resultados, predicciones y generalizaciones sobre la población estadística, basándose en la información contenida en las muestras representativas previamente elegidas por métodos de muestreo formales, se denomina *inferencia estadística*.

Es muy importante tener en cuenta que para medir el grado de representatividad de la muestra es necesario utilizar *muestreo probabilístico*. Diremos que el muestreo es probabilístico cuando pueda establecerse la probabilidad de obtener cada una de las muestras que sea posible seleccionar, esto es, cuando la selección de muestras constituya un fenómeno aleatorio probabilizable. Dicha selección se verificará en condiciones de azar, siendo susceptible de medida la incertidumbre derivada de la misma. Esto permitirá medir los errores cometidos en el proceso de muestreo (a través de varianza u otras medidas estadísticas).

Existen varios tipos de muestreo, dependiendo de que la población estadística sea finita o infinita, materia sobre la que existe amplia literatura estadística, pero nosotros consideraremos solamente el *muestreo en poblaciones finitas*. La población finita inicial que se desea investigar se denomina *población objetivo*, pero el muestreo de toda la población objetivo no siempre es posible debido a diferentes problemas que no permiten obtener información de algunos de sus elementos (inaccesibilidad de algunos de sus elementos, negativas a colaborar, ausencias, etc.), con lo que la población que realmente es objeto de estudio o *población investigada* no coincide con la población objetivo.

Por otro lado, para seleccionar la muestra, necesitaremos un listado de unidades de muestreo denominado *marco* que teóricamente debiera coincidir con la población objetivo. Un marco será más adecuado cuanto mejor cubra la población objetivo, es decir, cuanto menor sea el *error de cobertura*. Pero en los marcos son inevitables las desactualizaciones, las omisiones de algunas unidades, las duplicaciones de otras y la presencia de unidades extrañas y otras impurezas que obligan a su depuración (*depuración de marcos imperfectos*). Idealmente podría conseguirse la población objetivo eliminando del marco las unidades erróneamente incluidas en él (unidades extrañas, duplicaciones, etc.) y añadiendo las omisiones. Asimismo, también sería una meta que al eliminar del marco las unidades de las que no se puede obtener información (inaccesibles, ausentes, no colaboradoras, etc.) se obtuviera la población investigada. El marco puede estar constituido por unidades elementales de muestreo o por unidades compuestas. Una *unidad elemental (o simple)* es la unidad de muestreo más sencilla posible y una *unidad compuesta (o primaria)* está formada por varias unidades elementales. Como en la práctica no es fácil disponer de marcos de unidades elementales, se intenta conseguir marcos de unidades compuestas que son más accesibles. Por ejemplo, para estudiar habitantes de una región es más fácil disponer de un listado de hogares que de un listado de individuos. Se selecciona la muestra de un marco de hogares (unidades compuestas de varios individuos) y después se estudian las propiedades de los individuos con técnicas adecuadas.

MUESTREO Y ESTIMADORES. DISTRIBUCIONES EN EL MUESTREO

Consideramos los sucesos elementales asociados a un fenómeno o experimento aleatorio dado S_1, S_2, ..., S_m, entendiendo por *sucesos elementales* los más simples posibles, es decir, aquellos que no pueden ser descompuestos en otros sucesos. El conjunto $\{S_1, S_2, ..., S_m\}$ se denomina *espacio muestral* asociado al fenómeno o experimento. Si consideramos como fenómeno o experimento la extracción aleatoria de muestras dentro de una población por un procedimiento o método de muestreo dado, podemos considerar como sucesos elementales las muestras obtenidas, constituyendo el conjunto de las mismas el espacio muestral.

Habitualmente en los métodos de muestreo comunes se consideren iguales muestras con los mismos elementos, aunque estén colocados en orden diferente (el orden de colocación no interviene). Una muestra de tamaño n extraída de una población $U = \{U_1, U_2, ..., U_N\}$ de tamaño N mediante un método de muestreo dado, suele denotarse como $s = \{u_1, u_2, ..., u_n\}$. De esta forma, El conjunto de las N^n muestras posibles de tamaño n que se pueden formar con los N elementos de la población U es el espacio muestral S.

Evidentemente, para establecer la probabilidad de todas las muestras posibles derivadas de un procedimiento de muestreo dado, será necesario conocer ese conjunto de muestras; es decir, será necesario delimitar tanto el método de muestreo como el espacio muestral derivado del mismo. Un *procedimiento,* o *método, de muestreo* es sencillamente un proceso o mecanismo mediante el que se seleccionan las muestras de modo que cada una tenga una determinada probabilidad de ser elegida. Por tanto, el método aleatorio empleado para seleccionar la muestra define en el espacio muestral S una función de probabilidad P tal que:

- $P(S_i) \geq 0 \ \forall i$
- $\sum_S P(S_i) = 1$

A partir de una muestra, seleccionada mediante un determinado método de muestreo, se estiman las características poblacionales (media, total, proporción, etc.), con un error cuantificable y controlable. Las estimaciones se realizan a través de funciones matemáticas de la muestra denominadas *estimadores*, que se convierten en variables aleatorias al considerar la variabilidad de selección de las muestras. Los errores se cuantifican mediante varianzas, desviaciones típicas o errores cuadráticos medios de los estimadores, que miden la precisión de los mismos.

Para formalizar el problema de la estimación en poblaciones finitas, se considera que tenemos definida una característica X en la población U que toma el valor numérico X_i sobre la unidad U_i $i = 1, 2, ..., n$. Consideramos ahora una cierta función θ de los N valores X_i, por ejemplo, el total poblacional $\theta(X_1, ..., X_N) = \sum X_i$ para la característica X, o la media poblacional $\theta(X_1, ..., X_N) = (\sum X_i)/N$ para la característica X, que suele denominarse *parámetro poblacional*. Seleccionamos una muestra s, y a partir de ella queremos estimar el parámetro poblacional θ mediante una función $\hat{\theta} = \hat{\theta}(s(X)) = \hat{\theta}(X_1, ..., X_n)$, basada en los valores X_i $i = 1, 2, ..., n,$ que toma la característica X sobre las unidades de la muestra s (por ejemplo, el total muestral $\hat{\theta}(X_1, ..., X_n) = \sum X_i$, o la media muestral $\hat{\theta}(X_1, ..., X_n) = (\sum X_i)/n$, para estimar el total poblacional o la media poblacional, respectivamente. La función $\hat{\theta}$ que asocia a cada muestra s el valor numérico $\hat{\theta}(s(X)) = \hat{\theta}(X_1, ..., X_n)$, se denomina *estimador* del parámetro poblacional θ. A los valores $\hat{\theta}(s(X))$ para cada s, se los denomina *estimaciones*.

Dada la muestra $s = \{u_1, u_2, ..., u_n\}$, es habitual especificar el conjunto de valores X_i $i = 1, 2, ..., n$ que toma la característica X sobre las unidades de la muestra s mediante $s(X)=\{X_1, X_2, ..., X_n\}$. Al considerar todas las muestras s del espacio muestral S asociado al procedimiento de muestreo, y los valores que toma la característica X sobre dichas muestras, se obtiene el conjunto $S(X) = \{s(X) / s \in S\}$. Por tanto, podemos formalizar el concepto de estimador $\hat{\theta}$ para el parámetro poblacional θ definiéndolo mediante la aplicación medible:

$$\hat{\theta} : S(X) \subset R^n \rightarrow R$$
$$(X_1 \cdots X_n) \rightarrow \hat{\theta}(X_1 \cdots X_n) = t$$

Ya tenemos definido el estimador como un estadístico función de los valores que toma la característica X sobre los elementos del espacio muestral (muestras). Como ejemplos más sencillos de estimadores de los parámetros poblaciones total poblacional y media poblacional, tenemos los estimadores total muestral $\hat{X}$ y media muestral $\hat{\overline{X}}$, definidos como se indica a continuación:

$$\hat{\theta}_1 : S(X) \subset R^n \rightarrow R$$
$$(X_1 \cdots X_n) \rightarrow \hat{\theta}_1(X_1 \cdots X_n) = X_1 + \cdots + X_n = \hat{X}$$

$$\hat{\theta}_2 : S(X) \subset R^n \rightarrow R$$
$$(X_1 \cdots X_n) \rightarrow \hat{\theta}_2(X_1 \cdots X_n) = \frac{X_1 + \cdots + X_n}{n} = \hat{\overline{X}}$$

En cuanto a la construcción del estimador, ha de ser tal que la función $\hat{\theta}$ que asocia a cada muestra s el valor numérico $\hat{\theta}(s(X)) = \hat{\theta}(X_1, ..., X_n)$ sea calculable y esté definida para todas las muestras s del espacio muestral S generado por el procedimiento de muestreo considerado. La formación de estimadores no es una operación independiente del procedimiento de muestreo que se adopte. Generalmente, para construir estimadores se utiliza el *principio de analogía*; es decir, se estima un parámetro poblacional a partir del estimador muestral análogo. Por ejemplo, para estimar la media poblacional, la razón poblacional, etc., se utilizan como estimadores sus análogos muestrales, es decir, la media muestral, la razón muestral, etc. No siempre estos estimadores por analogía tienen las propiedades más deseables, pero suelen ser siempre consistentes, y a veces puede corregirse su sesgo multiplicándolos por una constante convenientemente elegida.

Distribución de un estimador en el muestreo

Se denomina distribución de probabilidad de una variable aleatoria a la función que asigna probabilidad a los valores que puede tomar la variable. Cuando se especifican los posibles valores de la variable aleatoria y sus probabilidades respectivas, tenemos construido el modelo de distribución de probabilidad. En nuestro caso la variable aleatoria es el estimador, y los posibles valores que puede tomar son las estimaciones, con lo que habremos obtenido la distribución de probabilidad en el muestreo para el estimador cuando conozcamos todos los valores posibles del estimador junto con las probabilidades de que el estimador tome cada valor.

En el párrafo anterior hemos formalizado el concepto de estimador $\hat{\theta}$ para el parámetro poblacional θ, definiéndolo mediante la variable aleatoria (aplicación medible):

$$\hat{\theta} : S(X) \subset R^n \rightarrow R$$
$$(X_1 \cdots X_n) \rightarrow \hat{\theta}(X_1 \cdots X_n) = t$$

Sea $T = \{t \in R / \exists (X_1, ..., X_n) \in S(X)$ que cumple $\hat{\theta}(X_1, ..., X_n) = t\}$. El conjunto $T \subset R$ constituye el conjunto de valores del estimador. Ahora vamos a definir las probabilidades de que el estimador tome estos valores (ley de probabilidad de la variable aleatoria $\hat{\theta}$) como sigue:

$$P^T(\hat{\theta}(X_1, ..., X_n) = t) = \sum_{\{S_i / \hat{\theta}(S_i(X))=t\}} P(s_i)$$

Al par $\{T, P^T\}$, formado por el conjunto de todos los posibles valores del estimador y por las probabilidades de que el estimador tome esos valores, se lo denomina *distribución del estimador en el muestreo*. A partir de la introducción del concepto de muestreo probabilístico y del conocimiento de la distribución de los estimadores en el muestreo, tanto la teoría de la probabilidad como la inferencia estadística están disponibles para ser aplicadas al muestreo. En todo el desarrollo de este libro se supone la existencia de muestreo probabilístico.

PROPIEDADES Y PRECISIÓN DE LOS ESTIMADORES. COMPARACIÓN DE ESTIMADORES

Como un estimador $\hat{\theta}$ de un parámetro poblacional θ es sencillamente una variable aleatoria unidimensional, nos interesarán sus características de centralización y dispersión, particularmente su esperanza, su varianza y sus momentos, así como otras medidas relativas a su precisión.

Precisión de los estimadores

Para analizar la precisión de un estimador suelen utilizarse los conceptos de error de muestreo (o desviación típica), acuracidad (o error cuadrático medio) y sesgo. Suele llamarse precisión a la acuracidad, lo que no es del todo correcto, ya que, aunque la acuracidad sea la magnitud más general para la medición de la precisión, hay casos en los que el análisis puede realizarse en función de otras magnitudes, como el sesgo o la desviación típica. Todas estas magnitudes que influyen en la precisión de un estimador pueden relacionarse a partir de la *descomposición del error cuadrático medio en sus componentes* de la forma siguiente:

$$ECM\left(\hat{\theta}\right) = \sigma\left(\hat{\theta}\right)^2 + B\left(\hat{\theta}\right)^2$$

Por tanto, la acuracidad (error cuadrático medio) de un estimador se descompone en la suma del cuadrado del error de muestreo y el cuadrado del sesgo.

En la práctica, se considera que el sesgo de $\hat{\theta}$ no es influyente cuando $\left|\frac{B\left(\hat{\theta}\right)}{\sigma\left(\hat{\theta}\right)}\right| < \frac{1}{10}$.

Comparación de estimadores insesgados

Un estimador $\hat{\theta}$ insesgado para el parámetro poblacional θ tiene la propiedad de que su error cuadrático medio coincide con su varianza, ya que al ser $E\left(\hat{\theta}\right) = \theta$ se tiene:

$$V\left(\hat{\theta}\right) = E\left(\hat{\theta} - E(\hat{\theta})\right)^2 = E\left(\hat{\theta} - \theta\right)^2 = ECM(\hat{\theta})$$

De esta forma los conceptos de acuracidad y error del estimador son similares para estimadores insesgados. Por tanto, *para comparar varios estimadores insesgados* $\hat{\theta}_i$ *del parámetro poblacional* θ en cuanto a precisión bastará considerar sus errores de muestreo $\sigma\left(\hat{\theta}_i\right) = +\sqrt{V\left(\hat{\theta}_i\right)}$, siendo más preciso el estimador que menor error de muestreo presente.

También en el caso de insesgadez el concepto de error relativo de muestreo puede expresarse en términos de una única magnitud variable $\sigma\left(\hat{\theta}\right)$ ya que:

$$CV\left(\hat{\theta}\right) = \frac{\sigma\left(\hat{\theta}\right)}{E(\hat{\theta})} = \frac{\sigma\left(\hat{\theta}\right)}{\theta}$$

y al ser θ una constante el error relativo está en función sólo del error de muestreo.

Con lo que resulta que, en el caso de estimadores insesgados, la precisión puede hacerse depender exclusivamente del error de muestreo $\sigma\left(\hat{\theta}\right)$.

Comparación de estimadores sesgados

Para estimadores $\hat{\theta}$ sesgados del parámetro poblacional θ, la magnitud general para analizar su precisión es su error cuadrático medio. Por tanto, para comparar varios estimadores sesgados del parámetro poblacional θ en cuanto a precisión se utilizará el error cuadrático medio y el estimador más preciso será el que menor error cuadrático medio presente.

Pero en la práctica el cálculo del error cuadrático medio puede ser problemático. Por esta razón, *cuando se intentan comparar varios estimadores* $\hat{\theta}_i$ *del parámetro poblacional* θ *todos sesgados*, se calcula para cada uno de ellos la cantidad:

$$\left| \frac{B\left(\hat{\theta}_i\right)}{\sigma\left(\hat{\theta}_i\right)} \right|$$

siendo más preciso aquel estimador que presenta una relación del sesgo al error de muestreo en valor absoluto más pequeña. También puede utilizarse el coeficiente de variación $CV\left(\hat{\theta}_i\right) = \sigma(\hat{\theta}_i)/E(\hat{\theta}_i)$, siendo más preciso el estimador con menor coeficiente de variación (error relativo). Se observa que el denominador del coeficiente de variación es el valor esperado del estimador, con lo que el coeficiente de variación recoge el efecto de un posible sesgo en el estimador.

Si los estimadores sesgados a comparar tienen todos sesgo despreciable, es decir, $\left|B(\hat{\theta}_i)/\sigma(\hat{\theta}_i)\right| < 1/10$, se compararían como si fuesen insesgados, de acuerdo con lo expresado en el apartado anterior.

Comparación de estimadores sesgados e insesgados

Para comparar en cuanto a precisión varios estimadores $\hat{\theta}_i$ unos sesgados y otros insesgados del parámetro poblacional θ, se utilizará el error cuadrático medio, y el estimador más preciso será el que menor error cuadrático medio presente. A veces, ante las dificultades de cálculo del error cuadrático medio se utiliza el coeficiente de variación $CV\left(\hat{\theta}_i\right) = \sigma(\hat{\theta}_i)/E(\hat{\theta}_i)$ (que contempla el posible efecto del sesgo en su denominador), siendo más preciso el estimador con menor coeficiente de variación (error relativo).

Si los estimadores sesgados tienen todos sesgo despreciable, $\left|B(\hat{\theta}_i)/\sigma(\hat{\theta}_i)\right| < 1/10$, se haría la comparación global como insesgados de acuerdo con los valores de $\sigma(\hat{\theta}_i)$.

Cuantificación de la ganancia en precisión de los estimadores

Para medir la precisión de los estimadores suele utilizarse el error cuadrático medio, el error relativo (coeficiente de variación) o el error de muestreo (desviación típica). En cada caso, la ganancia en precisión estará dada por las respectivas tasas de variación:

$$\left(\frac{ECM(\hat{\theta}_1)}{ECM(\hat{\theta}_2)}-1\right)x100 \qquad \left(\frac{CV(\hat{\theta}_1)}{CV(\hat{\theta}_2)}-1\right)x100 \qquad \left(\frac{\sigma(\hat{\theta}_1)}{\sigma(\hat{\theta}_2)}-1\right)x100$$

ESTIMACIÓN POR INTERVALOS DE CONFIANZA

Al estimar parámetros de la población en estudio basándose en la información contenida en la muestra, pueden usarse los valores puntuales de un estadístico basado en la misma, o puede utilizarse un intervalo de valores dentro del cual se tiene confianza de que esté el valor del parámetro. En el primer caso estamos ante el proceso de *estimación puntual,* en el que utilizamos directamente los valores de un estadístico, denominado *estimador puntual,* sobre la muestra dada (*estimaciones puntuales*), para estimar los valores poblacionales. En el segundo caso estamos ante la *estimación por intervalos*, donde se calcula un intervalo de confianza en el que razonablemente cae el valor estimado con un *nivel de confianza* prefijado.

Obtener una estimación por intervalos (o definir un intervalo de confianza) para un parámetro poblacional θ al nivel de confianza α consiste en hallar un intervalo real para el que se tiene una probabilidad $1-\alpha$ de que el verdadero valor del parámetro θ caiga dentro del citado intervalo. El valor $1-\alpha$ suele denominarse *coeficiente de confianza*.

Intervalos de confianza cuando el estimador es insesgado

En este caso se persigue estimar el parámetro poblacional θ mediante un intervalo de confianza basado en el estimador $\hat{\theta}$ insesgado para θ ($E(\hat{\theta}) = \theta$). Para estimadores insesgados, es necesario distinguir entre el caso en que la distribución del estimador es normal y el caso en que dicha distribución no puede asegurarse que sea normal.

a) El estimador $\hat{\theta}$ tiene una distribución normal

El intervalo de confianza para el parámetro poblacional θ basado en $\hat{\theta}$ será:

$$\left[\hat{\theta}-\lambda_\alpha\sigma\left(\hat{\theta}\right),\hat{\theta}+\lambda_\alpha\sigma\left(\hat{\theta}\right)\right] \text{ con } \lambda_\alpha = F^{-1}_{N(0,1)}\left(1-\frac{\alpha}{2}\right)$$

F es la función de distribución de la normal (0,1), y α es el nivel de confianza. Si realmente es dudoso que $\hat{\theta}$ tenga una distribución normal, puede utilizarse la distribución t de Student con $n-1$ grados de libertad para calcular el intervalo de confianza para θ que, en este caso, será:

$$\left[\hat{\theta} - t_\alpha \hat{\sigma}\left(\hat{\theta}\right), \hat{\theta} + t_\alpha \hat{\sigma}\left(\hat{\theta}\right)\right] \text{ con } \; t_\alpha = F_{t_{n-1}}^{-1}\left(1 - \frac{\alpha}{2}\right)$$

F es la función de distribución de una *t* de Student con $n - 1$ grados de libertad.

b) El estimador $\hat{\theta}$ no tiene una distribución normal

El intervalo de confianza, derivado de la desigualdad de Tchevichev, para el parámetro poblacional θ basado en $\hat{\theta}$ que cubre el valor de θ con una probabilidad $1 - \alpha$ (coeficiente de confianza), será:

$$\left[\hat{\theta} - \frac{\sigma\left(\hat{\theta}\right)}{\sqrt{\alpha}}, \hat{\theta} + \frac{\sigma\left(\hat{\theta}\right)}{\sqrt{\alpha}}\right]$$

Este intervalo suele ser más ancho que el obtenido cuando la distribución de $\hat{\theta}$ es normal. A medida que $\hat{\theta}$ se aleja más de la normalidad, la anchura de este intervalo es mucho mayor respecto del obtenido para normalidad. Ya sabemos que una estimación por intervalos es tanto mejor cuanto más reducido sea el intervalo de confianza correspondiente; de ahí que la propiedad de normalidad sea muy deseable, pues en este caso los intervalos obtenidos son muy estrechos, lo que implica una buena estimación por intervalos.

Intervalos de confianza en estimadores sesgados

El intervalo de confianza para θ basado en el estimador $\hat{\theta}$ en presencia del sesgo no despreciable $B\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$ es el siguiente:

$$\left[\hat{\theta} - \lambda_\alpha \sigma\left(\hat{\theta}\right) - \left| B\left(\hat{\theta}\right)\right|, \hat{\theta} + \lambda_\alpha \sigma\left(\hat{\theta}\right) - \left| B\left(\hat{\theta}\right)\right|\right]$$

Observamos que se trata de un intervalo no centrado en $\hat{\theta}$ y desplazado en la cantidad $B\left(\hat{\theta}\right)$ respecto del intervalo sin sesgo, que debe centrarse situándonos en la peor de las circunstancias, es decir, tomando como extremo fijo del intervalo el más lejano del centro $\hat{\theta}$, y calculando el otro extremo por equidistancia al centro. Ante esta situación, la presencia del sesgo $B\left(\hat{\theta}\right)$ origina que el intervalo de confianza para θ basado en el estimador $\hat{\theta}$ y centrado en $\hat{\theta}$, tenga una longitud superior al intervalo cuando no hay sesgo. Por tanto, la presencia de sesgo conduce a una estimación por intervalos menos precisa.

El intervalo de confinaza ya centrado será el siguiene:

$$\left[\hat{\theta} - \lambda_\alpha \sigma\left(\hat{\theta}\right) - \left| B\left(\hat{\theta}\right)\right|, \hat{\theta} + \lambda_\alpha \sigma\left(\hat{\theta}\right) + \left| B\left(\hat{\theta}\right)\right|\right]$$

PROBLEMAS RESUELTOS

1.1. Sobre las regiones que componen un determinado país se mide la variable X=Número de personas activas, obteniendo como resultados 6 millones, 4 millones, 3 millones y 8 millones con probabilidades iniciales de selección 1/6, 1/3, 1/3 y 1/6, respectivamente, para cada región. Se trata de estimar en millones de personas la cifra media de actividad, extrayendo muestras de la variable X con tamaño 2 sin reposición y sin tener en cuenta el orden de colocación de sus elementos. Para ello se consideran los estimadores alternativos MEDIANA y MEDIA ARMÓNICA. Se pide lo siguiente:

1) Especificar el espacio muestral definido por este procedimiento de muestreo, las probabilidades asociadas a las muestras y la distribución en el muestreo de los dos estimadores. Analizar la precisión de los dos estimadores. ¿Cuál de ellos es mejor?

2) Hallar intervalos de confianza para la mediana y la media armónica basados en la muestra de mayor probabilidad para un nivel de confianza del 2 por mil (α = 0,002). Como dato se sabe que $F^{-1}(0.999)= 3$, siendo F la función de distribución de la normal (0,1). Comentar los resultados relacionándolos con los del apartado 1.

Tenemos un procedimiento de muestreo sin reposición en el que no interviene el orden de colocación de las unidades en las muestras, con lo que el espacio muestral tendrá $\binom{4}{2} = 6$ muestras.

A continuación se especifican las muestras, sus probabilidades y los valores de los estimadores mediana $\hat{M}$ y media armónica $\hat{\bar{X}}_H$ para cada muestra.

$S(X)$	$P(X)$	$\hat{M}$	$\hat{\bar{X}}_H$
(6 4)	3/20	5	24/5
(6 3)	3/20	9/2	4
(6 8)	1/15	7	48/7
(4 3)	1/3	7/2	24/7
(4 8)	3/20	6	16/3
(3 8)	3/20	11/2	48/11

Dado que no hay reposición y que no importa el orden de colocación de los elementos en las muestras (muestras con los mismos elementos colocados en orden diferente se consideran la misma muestra), las probabilidades de la columna P(X) se han calculado de la siguiente forma:

$$P(6,4) = P\{6,4\} + P\{4,6\} = P(6)P(4/6) + P(4)P(6/4) = \frac{1}{6}\cdot\frac{2}{5} + \frac{2}{6}\cdot\frac{1}{4} = \frac{3}{20}$$

$$P(6,3) = P\{6,3\} + P\{3,6\} = P(6)P(3/6) + P(3)P(6/3) = \frac{1}{6}\cdot\frac{2}{5} + \frac{2}{6}\cdot\frac{1}{4} = \frac{3}{20}$$

$$P(6,8) = P\{6,8\} + P\{8,6\} = P(6)P(8/6) + P(8)P(6/8) = \frac{1}{6}\cdot\frac{1}{5} + \frac{1}{6}\cdot\frac{1}{5} = \frac{1}{15}$$

$$P(4,3) = P\{4,3\} + P\{3,4\} = P(4)P(3/4) + P(3)P(4/3) = \frac{2}{6}\cdot\frac{2}{4} + \frac{2}{6}\cdot\frac{2}{4} = \frac{1}{3}$$

$$P(4,8) = P\{4,8\} + P\{8,4\} = P(4)P(8/4) + P(8)P(4/8) = \frac{2}{6}\cdot\frac{1}{4} + \frac{1}{6}\cdot\frac{2}{5} = \frac{3}{20}$$

$$P(3,8) = P\{3,8\} + P\{8,3\} = P(3)P(8/3) + P(8)P(3/8) = \frac{2}{6}\cdot\frac{1}{4} + \frac{1}{6}\cdot\frac{2}{5} = \frac{3}{20}$$

Las probabilidades anteriores también pueden calcularse mediante la expresión $P(u_i,u_j) = P(u_i)P(u_j/u_i) + P(u_j)P(u_i/u_j) = P(u_i)P(u_j)/(1-P(u_i)) + P(u_j)P(u_i)/(1-P(u_j)) = PiPj/(1-Pi) + PiPj(1-Pj)$.

Las distribuciones de probabilidad de los dos estimadores se calcularán mediante la expresión ya conocida $P^T(\hat{\theta}(X_1, ..., X_n) = t) = \sum_{\{S_i / \hat{\theta}(S_i(X))=t\}} P(S_i)$, de la siguiente forma:

$$\hat{M}\begin{cases} P^T(\hat{M}=5) = P(6,4) = \frac{3}{20} \\ P^T(\hat{M}=\frac{9}{2}) = P(6,3) = \frac{3}{20} \\ P^T(\hat{M}=7) = P(6,8) = \frac{1}{15} \\ P^T(\hat{M}=\frac{7}{2}) = P(4,3) = \frac{1}{3} \\ P^T(\hat{M}=6) = P(4,8) = \frac{3}{20} \\ P^T(\hat{M}=\frac{11}{2}) = P(3,8) = \frac{3}{20} \end{cases} \qquad \hat{\bar{X}}_H\begin{cases} P^T(\hat{\bar{X}}_H=\frac{24}{5}) = P(6,4) = \frac{3}{20} \\ P^T(\hat{\bar{X}}_H=4) = P(6,3) = \frac{3}{20} \\ P^T(\hat{\bar{X}}_H=\frac{48}{7}) = P(6,8) = \frac{1}{15} \\ P^T(\hat{\bar{X}}_H=\frac{24}{7}) = P(4,3) = \frac{1}{3} \\ P^T(\hat{\bar{X}}_H=\frac{16}{3}) = P(4,8) = \frac{3}{20} \\ P^T(\hat{\bar{X}}_H=\frac{48}{11}) = P(3,8) = \frac{3}{20} \end{cases}$$

Una vez conocida la distribución de probabilidad en el muestreo de los dos estimadores analizaremos si son insesgados o no. Para ello calculamos en primer lugar los valores de la mediana y media armónica poblacionales como sigue:

$$M = (4+6)/2 = 5 \qquad \bar{X}_H = \frac{4}{1/6+1/4+1/3+1/8} = 4{,}57$$

Ahora, para comprobar la insesgadez, hallamos la esperanza de los estimadores:

$$E(\hat{M}) = 5\cdot\frac{3}{20} + \frac{9}{2}\cdot\frac{3}{20} + 7\cdot\frac{1}{15} + \frac{7}{2}\cdot\frac{1}{3} + 6\cdot\frac{3}{20} + \frac{11}{2}\cdot\frac{3}{20} = 4{,}78 \neq \bar{X} = 5$$

$$E(\hat{\bar{X}}_H) = \frac{24}{5}\cdot\frac{3}{20} + 4\cdot\frac{3}{20} + \frac{48}{7}\cdot\frac{1}{15} + \frac{24}{7}\cdot\frac{1}{3} + \frac{16}{3}\cdot\frac{3}{20} + \frac{48}{11}\cdot\frac{3}{20} = 4{,}37 \neq \bar{X}_H = 4{,}57$$

Vemos que los dos estimadores son sesgados y los valores de sus sesgos son:

$$B(\hat{M}) = E(\hat{\bar{X}}) - \bar{X} = 4{,}78 - 5 = -0{,}22 \qquad B(\hat{\bar{X}}_H) = E(\hat{\bar{X}}_H) - \bar{X}_H = 4{,}37 - 4{,}57 = -0{,}2$$

Ahora calculamos las varianzas de los dos estimadores como sigue:

$$V(\hat{M}) = E(\hat{M} - 4{,}78)^2 = (5-4{,}78)^2\cdot\frac{3}{20} + (\frac{9}{2}-4{,}78)^2\cdot\frac{3}{20} + (7-4{,}78)^2\cdot\frac{1}{15} + (\frac{7}{2}-4{,}78)^2\cdot\frac{1}{3} + (6-4{,}78)^2\cdot\frac{3}{20} + (\frac{11}{2}-4{,}78)^2\cdot\frac{3}{20} = 1{,}19$$

$$V(\hat{\bar{X}}_H) = E(\hat{\bar{X}}_H - 4{,}37)^2 = (\frac{24}{5} - 4{,}37)^2 \cdot \frac{3}{20} + (4 - 4{,}37)^2 \cdot \frac{3}{20} + (\frac{48}{7} - 4{,}37)^2 \cdot \frac{1}{15}$$
$$+ (\frac{24}{7} - 4{,}37)^2 \cdot \frac{1}{3} + (\frac{16}{3} - 4{,}37)^2 \cdot \frac{3}{20} + \frac{48}{11} \cdot (\frac{3}{20} - 4{,}37)^2 = 0{,}89$$

Ya que los dos estimadores son sesgados se pueden hacer las comparaciones a través del error cuadrático medio, pero antes se deben calcular las cantidades $\left|\frac{B(\hat{\theta}_i)}{\sigma(\hat{\theta}_i)}\right|$ para ver si el sesgo es o no despreciable. Tenemos:

$$\left|\frac{B(\hat{M})}{\sigma(\hat{M})}\right| = \frac{0{,}22}{\sqrt{1{,}19}} = 0.2, \quad \left|\frac{B(\hat{\bar{X}}_H)}{\sigma(\hat{\bar{X}}_H)}\right| = \frac{0{,}2}{\sqrt{0{,}89}} = 0{,}2$$

Los dos valores son superiores a 1/10, con lo que el sesgo no resulta despreciable en ningún caso (los dos estimadores son igualmente precisos según la razón del sesgo a la desviación típica). Calculamos ahora los errores cuadráticos medios para aquilatar mejor la diferencia de precisiones y ver realmente qué estimador es mejor.

$$ECM(\hat{M}) = E(\hat{M} - 5)^2 = (5-5)^2 \cdot \frac{3}{20} + (\frac{9}{2} - 5)^2 \cdot \frac{3}{20} + (7-5)^2 \cdot \frac{1}{15} + (\frac{7}{2} - 5)^2 \cdot \frac{1}{3} + (6-5)^2 \cdot \frac{3}{20} + (\frac{11}{2} - 5)^2 \cdot \frac{3}{20} = 1{,}24$$

$$ECM(\hat{\bar{X}}_H) = E(\hat{\bar{X}}_H - 4{,}57)^2 = (\frac{24}{5} - 4{,}57)^2 \cdot \frac{3}{20} + (4 - 4{,}57)^2 \cdot \frac{3}{20} + (\frac{48}{7} - 4{,}57)^2 \cdot \frac{1}{15} + (\frac{24}{7} - 4{,}57)^2 \cdot \frac{1}{3}$$
$$+ (\frac{16}{3} - 4{,}57)^2 \cdot \frac{3}{20} + \frac{48}{11} \cdot (\frac{3}{20} - 4{,}57)^2 = 0{,}93$$

El mejor estimador resulta ser la media armónica porque tiene menor error cuadrático medio. Para cuantificar las ganancias en precisión calculamos:

$$\left(\frac{1{,}24}{0{,}93} - 1\right) \cdot 100 = 33{,}33$$

Se observa que el uso de la media armónica mejora en un 33,33% la estimación a partir de la mediana.

Para calcular los intervalos de confianza par la mediana y la media armónica basados en la muestra de mayor probabilidad (4,3), una vez que ya sabemos que son sesgados con sesgo influyente (no despreciable), utilizamos la expresión:

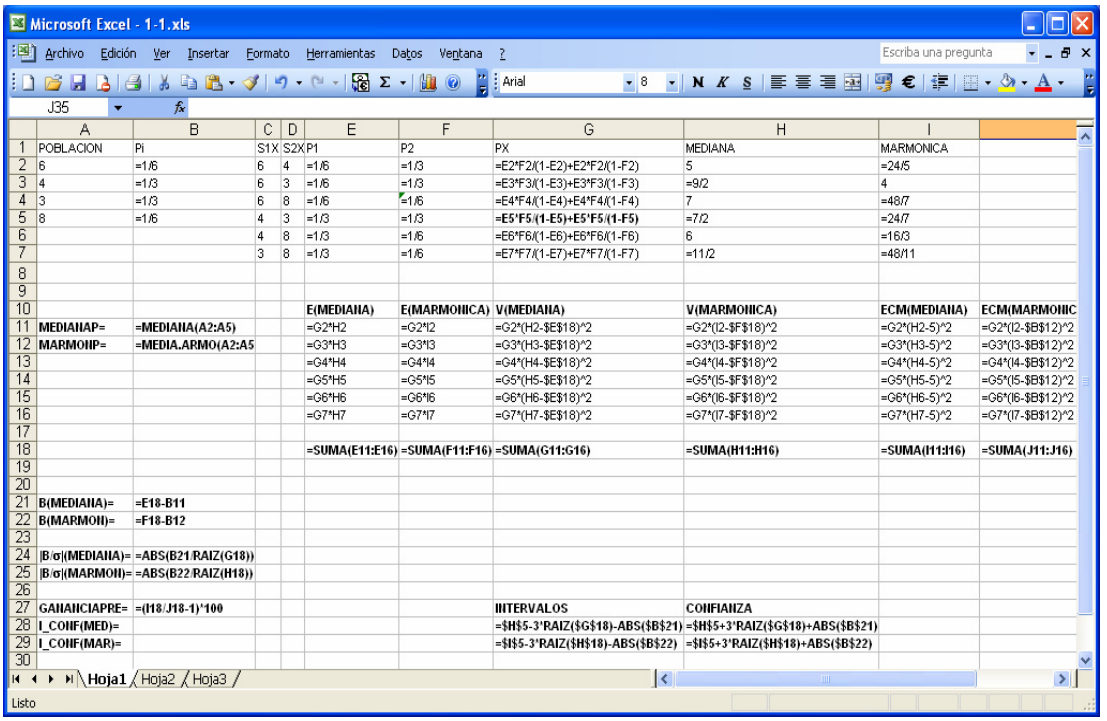
$$[\hat{\theta} - \lambda_\alpha \sigma(\hat{\theta}) - |B(\hat{\theta})|, \hat{\theta} + \lambda_\alpha \sigma(\hat{\theta}) + |B(\hat{\theta})|]$$

Tenemos:

$$\hat{M} \rightarrow [7/2 - 3\sqrt{1{,}19} - |-0{,}22|, \quad 7/2 + 3\sqrt{1{,}19} + |-0{,}22|] = [0{,}004, \quad 6{,}99]$$
$$\hat{X}_H \rightarrow [24/7 - 3\sqrt{0{,}89} - |-0{,}2|, \quad 24/7 + 3\sqrt{0{,}89} + |-0{,}2|] = [0{,}39, \quad 6{,}45]$$

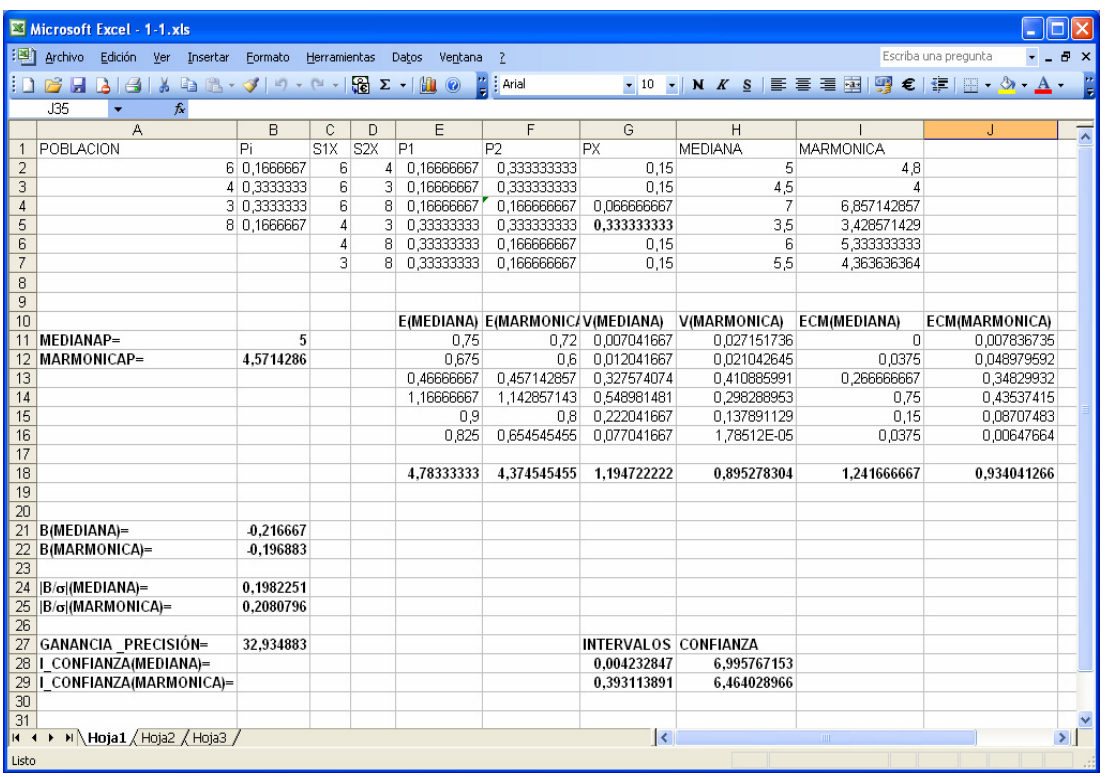
Se observa que el intervalo más estrecho es el relativo a la media armónica, ya que es el estimador más preciso.

Los cálculos pueden automatizarse con EXCEL como sigue:



	A	B	C	D	E	F	G	H	I	
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIANA	MARMONICA	
2	6	=1/6	6	4	=1/6	=1/3	=E2*F2/(1-E2)+E2*F2/(1-F2)	5	=24/5	
3	4	=1/3	6	3	=1/6	=1/3	=E3*F3/(1-E3)+E3*F3/(1-F3)	=9/2	4	
4	3	=1/3	6	8	=1/6	=1/6	=E4*F4/(1-E4)+E4*F4/(1-F4)	7	=48/7	
5	8	=1/6	4	3	=1/3	=1/3	=E5*F5/(1-E5)+E5*F5/(1-F5)	=7/2	=24/7	
6			4	8	=1/3	=1/6	=E6*F6/(1-E6)+E6*F6/(1-F6)	6	=16/3	
7			3	8	=1/3	=1/6	=E7*F7/(1-E7)+E7*F7/(1-F7)	=11/2	=48/11	
8										
9										
10					E(MEDIANA)	E(MARMONICA)	V(MEDIANA)	V(MARMONICA)	ECM(MEDIANA)	ECM(MARMONIC
11	MEDIANAP=	=MEDIANA(A2:A5)			=G2*H2	=G2*I2	=G2*(H2-$E$18)^2	=G2*(I2-$F$18)^2	=G2*(H2-5)^2	=G2*(I2-$B$12)^2
12	MARMONP=	=MEDIA.ARMO(A2:A5			=G3*H3	=G3*I3	=G3*(H3-$E$18)^2	=G3*(I3-$F$18)^2	=G3*(H3-5)^2	=G3*(I3-$B$12)^2
13					=G4*H4	=G4*I4	=G4*(H4-$E$18)^2	=G4*(I4-$F$18)^2	=G4*(H4-5)^2	=G4*(I4-$B$12)^2
14					=G5*H5	=G5*I5	=G5*(H5-$E$18)^2	=G5*(I5-$F$18)^2	=G5*(H5-5)^2	=G5*(I5-$B$12)^2
15					=G6*H6	=G6*I6	=G6*(H6-$E$18)^2	=G6*(I6-$F$18)^2	=G6*(H6-5)^2	=G6*(I6-$B$12)^2
16					=G7*H7	=G7*I7	=G7*(H7-$E$18)^2	=G7*(I7-$F$18)^2	=G7*(H7-5)^2	=G7*(I7-$B$12)^2
17										
18					=SUMA(E11:E16)	=SUMA(F11:F16)	=SUMA(G11:G16)	=SUMA(H11:H16)	=SUMA(I11:I16)	=SUMA(J11:J16)
19										
20										
21	B(MEDIANA)=	=E18-B11								
22	B(MARMON)=	=F18-B12								
23										
24	|B/σ|(MEDIANA)=	=ABS(B21/RAIZ(G18))								
25	|B/σ|(MARMON)=	=ABS(B22/RAIZ(H18))								
26										
27	GANANCIAPRE=	=(I18/J18-1)*100					INTERVALOS	CONFIANZA		
28	I_CONF(MED)=						=$H$5-3*RAIZ($G$18)-ABS($B$21)	=$H$5+3*RAIZ($G$18)+ABS($B$21)		
29	I_CONF(MAR)=						=$I$5-3*RAIZ($H$18)-ABS($B$22)	=$I$5+3*RAIZ($H$18)+ABS($B$22)		
30										

Los resultados son los siguientes:



	A	B	C	D	E	F	G	H	I	J
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIANA	MARMONICA	
2	6	0,1666667	6	4	0,16666667	0,333333333	0,15	5	4,8	
3	4	0,3333333	6	3	0,16666667	0,333333333	0,15	4,5	4	
4	3	0,3333333	6	8	0,16666667	0,166666667	0,066666667	7	6,857142857	
5	8	0,1666667	4	3	0,33333333	0,333333333	**0,333333333**	3,5	3,428571429	
6			4	8	0,33333333	0,166666667	0,15	6	5,333333333	
7			3	8	0,33333333	0,166666667	0,15	5,5	4,363636364	
8										
9										
10					E(MEDIANA)	E(MARMONICA	V(MEDIANA)	V(MARMONICA)	ECM(MEDIANA)	ECM(MARMONICA)
11	MEDIANAP=	5			0,75	0,72	0,007041667	0,027151736	0	0,007836735
12	MARMONICAP=	4,5714286			0,675	0,6	0,012041667	0,021042645	0,0375	0,048979592
13					0,46666667	0,457142857	0,327574074	0,410885991	0,266666667	0,34829932
14					1,16666667	1,142857143	0,548981481	0,298288953	0,75	0,43537415
15					0,9	0,8	0,222041667	0,137891129	0,15	0,08707483
16					0,825	0,654545455	0,077041667	1,78512E-05	0,0375	0,00647664
17										
18					**4,78333333**	**4,374545455**	**1,194722222**	**0,895278304**	**1,241666667**	**0,934041266**
19										
20										
21	B(MEDIANA)=	-0,216667								
22	B(MARMONICA)=	-0,196883								
23										
24	|B/σ|(MEDIANA)=	0,1982251								
25	|B/σ|(MARMONICA)=	0,2080796								
26										
27	GANANCIA_PRECISIÓN=	32,934883					INTERVALOS	CONFIANZA		
28	I_CONFIANZA(MEDIANA)=						0,004232847	6,995767153		
29	I_CONFIANZA(MARMONICA)=						0,393113891	6,464028966		
30										
31										

1.2. Dada la población $\{U_1, U_2, U_3, U_4, U_5\}$ seleccionamos muestras de tamaño 3 por el siguiente método de muestreo: De un recipiente que contiene tres bolas numeradas del 1 al 3 se extraen al azar dos bolas mediante muestreo aleatorio sin reposición con probabilidades iguales, y a continuación, de otro recipiente con dos bolas numeradas con el 4 y el 5 se extrae una bola. Se supone que extraer la bola i-ésima equivale a elegir para la muestra la unidad U_i. Consideramos los estimadores por analogía siguientes:

$\hat{T}_1$ = Proporción de subíndices pares en la muestra

$\hat{T}_2$ = Total de subíndices impares en la muestra

1) Hallar las distribuciones en el muestreo de $\hat{T}_1$ y $\hat{T}_2$ y sus varianzas, sesgos y errores cuadráticos medios.

2) Comparar las precisiones de los estimadores anteriores cuantificando las ganancias en precisión tanto por la vía de la estimación puntual como por la vía de la estimación por intervalos al 95%. Comentar los resultados.

Para hallar el espacio muestral asociado a este procedimiento de muestreo consideramos la urna U_1 con tres bolas y la urna U_2 con dos bolas.

Como en la urna U_1 seleccionamos dos bolas sin reposición, las posibilidades son $(A_1\ A_2)$, $(A_1\ A_3)$ y $(A_2\ A_3)$. Como para cada par de bolas seleccionadas de la urna U_1 se selecciona una bola en la urna U_2, las posibles muestras de tres elementos serán $(A_1\ A_2\ A_4)$, $(A_1\ A_2\ A_5)$, $(A_1\ A_3\ A_4)$, $(A_1\ A_3\ A_5)$, $(A_2\ A_3\ A_4)$ y $(A_2\ A_3\ A_5)$.

Las probabilidades de las muestras se calculan como se indica a continuación:

$P(A_1\ A_2\ A_4) = P(A_1\ A_2\ /U_1)P(A_4/U_2) + P(A_2\ A_1/U_1)P(A_4/U_2) = P_1(A_1)P_1(A_2/A_1)P_2(A_4) + P_1(A_2\ )P_1(A_1\ /\ A_2)P_2(A_4) = (1/3)(1/2)(1/2) + (1/3)(1/2)(1/2) = 1/6$

$P(A_1\ A_2\ A_5) = P(A_1\ A_2/U_1)P(A_5/U_2) + P(A_2\ A_1/U_1)P(A_5/U_2) = P_1(A_1\ )P_1(A_2/A_1)P_2(A_5) + P_1(A_2)P_1(A_1/A_2)P_2(A_5) = (1/3)(1/2)(1/2) + (1/3)(1/2)(1/2) = 1/6$

El cálculo de las probabilidades de las restantes muestras es similar, y el valor es 1/6 para todas ellas; es decir, estamos ante un método de selección con probabilidades iguales. Ya podemos formar la tabla con las muestras del espacio muestral *S_X*, sus probabilidades P_i y los valores de los dos estimadores del problema sobre las mismas $\hat{T}_1$ y $\hat{T}_2$, datos que van a permitirnos el cálculo de las distribuciones en el muestreo de los estimadores. En el siguiente cuadro se especifican las muestras, sus probabilidades y los valores de los estimadores para cada muestra.

S_X	Pi	$\hat{T}_1$	$\hat{T}_2$
$A_1A_2A_4$	1 / 6	2 / 3	1
$A_1A_2A_5$	1 / 6	1 / 3	2
$A_1A_3A_4$	1 / 6	1 / 3	2
$A_1A_3A_5$	1 / 6	0	3
$A_2A_3A_4$	1 / 6	2 / 3	1
$A_2A_3A_5$	1 / 6	1 / 3	2

Las distribuciones de probabilidad de los dos estimadores se calcularán mediante la expresión ya conocida $P^T(\hat{\theta}(X_1, ...,X_n) = t) = \sum_{\{S_i / \hat{\theta}(S_i(X))=t\}} P(S_i)$, de la siguiente forma:

$$\hat{T}_1\begin{cases} P^T(\hat{T}_1 = 2/3) = 2\cdot\frac{1}{6} = \frac{1}{3} \\ P^T(\hat{T}_1 = 1/3) = 3\cdot\frac{1}{6} = \frac{1}{2} \\ P^T(\hat{T}_1 = 0) = \frac{1}{6} \end{cases} \qquad \hat{T}_1\begin{cases} P^T(\hat{T}_2 = 1) = 2\cdot\frac{1}{6} = \frac{1}{3} \\ P^T(\hat{T}_2 = 2) = 3\cdot\frac{1}{6} = \frac{1}{2} \\ P^T(\hat{T}_2 = 3) = \frac{1}{6} \end{cases}$$

Una vez conocida la distribución de probabilidad en el muestreo de los dos estimadores analizaremos si son insesgados o no. Para ello calculamos en primer lugar los valores de la proporción de subíndices pares de la población θ_1 y del total de subíndices impares de la población θ_2, que son los parámetros que estamos estimando con los estimadores $\hat{T}_1$ y $\hat{T}_2$, respectivamente. Se tiene:

$$\theta_1 = 2/3 \quad \theta_2 = 3$$

Ahora, para comprobar la insesgadez, hallamos la esperanza matemática de los estimadores tal y como se indica a continuación:

$$E(\hat{T}_1) = \frac{2}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{1}{2} + 0\cdot\frac{1}{6} = 7/18 = 0{,}388888888 \neq 2/5 = \theta_1$$

$$E(\hat{T}_2) = 1\cdot\frac{1}{3} + 2\cdot\frac{1}{2} + 3\cdot\frac{1}{6} = 11/6 = 1{,}8333333333 \neq 3 = \theta_2$$

El estimador $\hat{T}_1$ es sesgado con sesgo $B(\hat{T}_1) = E(\hat{T}_1) - \theta_1 = 7/18 - 2/5 = -1/90 = -0{,}0111$, y el estimador $\hat{T}_2$ también es sesgado con sesgo $B(\hat{T}_2) = E(\hat{T}_2) - \theta_2 = 11/6 - 3 = -7/6 = -1{,}16666$. Las varianzas de los estimadores son:

$$V(\hat{T}_1) = (\frac{2}{3} - 0{,}388)^2\cdot\frac{1}{3} + (\frac{1}{3} - 0{,}388)^2\cdot\frac{1}{2} + (0 - 0{,}388)^2\cdot\frac{1}{6} = 0{,}0524$$

$$V(\hat{T}_2) = (1 - 1{,}833)^2\cdot\frac{1}{3} + (2 - 1{,}833)^2\cdot\frac{1}{2} + (3 - 1{,}833)^2\cdot\frac{1}{6} = 0{,}4722$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(\hat{T}_1) = \sqrt{0{,}0524} = 0{,}2289 \quad y \quad \sigma(\hat{T}_2) = \sqrt{0{,}4722} = 0{,}687$$

Como $|B(\hat{T}_1)/\sigma(\hat{T}_1)| = 0{,}0485 < 1/10$, el sesgo del estimador $\hat{T}_1$ es despreciable, por lo que este puede considerarse a todos los efectos insesgado. Como $|B(\hat{T}_2)/\sigma(\hat{T}_2)| = 1{,}69 > 1/10$ el sesgo del estimador $\hat{T}_2$ no es despreciable, y como un estimador es sesgado y el otro insesgado, la comparación de estimadores puede hacerse a través de los errores cuadráticos medios. Tenemos:

$$ECM(\hat{T}_1) = (\frac{2}{3} - 0{,}4)^2 \cdot \frac{1}{3} + (\frac{1}{3} - 0{,}4)^2 \cdot \frac{1}{2} + (0 - 0{,}4)^2 \cdot \frac{1}{6} = 0{,}0526$$

$$ECM(\hat{T}_2) = (1 - 3)^2 \cdot \frac{1}{3} + (2 - 3)^2 \cdot \frac{1}{2} + (3 - 3)^2 \cdot \frac{1}{6} = 1{,}833$$

Se observa que el error cuadrático medio de $\hat{T}_1$ es prácticamente igual que su varianza, dado que es prácticamente insesgado. Evidentemente el mejor estimador es $\hat{T}_1$, pues su error cuadrático medio es mucho menor que el de $\hat{T}_2$. La ganancia en precisión por usar $\hat{T}_1$ en vez de $\hat{T}_2$ es:

$$GP = (EMC(\hat{T}_2)/EMC(\hat{T}_1) - 1)*100 = (1{,}833/0{,}0526 - 1)*100 = 3385{,}9\%$$

Para hallar un intervalo de confianza para $\hat{T}_1$ (que es insesgado) basado en la primera muestra y suponiendo normalidad en la población se utilizará la fórmula:

$$[\hat{T}_1 - \lambda_\alpha \sigma(\hat{T}_1), \hat{T}_1 + \lambda_\alpha \sigma(\hat{T}_1)] = [2/3 - 1.96*0.229,\ 2/3 + 1.96*0.229] = [0.217, 1.15]$$

Para el resto de las muestras se realizan cálculos similares.

Se puede suponer normalidad en la población porque el coeficiente de asimetría g_1 y el coeficiente de curtosis g_2 de $\hat{T}_1$ caen en el intervalo [-2,2]. El coeficiente de asimetría depende del momento de tercer orden centrado en la media m3 y el coeficiente de curtosis depende del momento de orden 4 centrado en la media m4 y se calculan como sigue:

$$g_1 = m3/\sigma^3 = -0{,}0027/0{,}229^3 = 0{,}22$$
$$g_2 = m4/\sigma^4 - 3 = -0{,}0058/0{,}229^4 - 3 = -0{,}89$$

$$m3(\hat{T}_1) = (\frac{2}{3} - 0{,}388)^3 \cdot \frac{1}{3} + (\frac{1}{3} - 0{,}388)^3 \cdot \frac{1}{2} + (0 - 0{,}388)^3 \cdot \frac{1}{6} = 0{,}22$$

$$m4(\hat{T}_1) = (\frac{2}{3} - 0{,}388)^4 \cdot \frac{1}{3} + (\frac{1}{3} - 0{,}388)^4 \cdot \frac{1}{2} + (0 - 0{,}388)^4 \cdot \frac{1}{6} = -0{,}89$$

Si no hubiera habido normalidad, el intervalo de confianza para $\hat{T}_1$ se habría calculado como sigue:

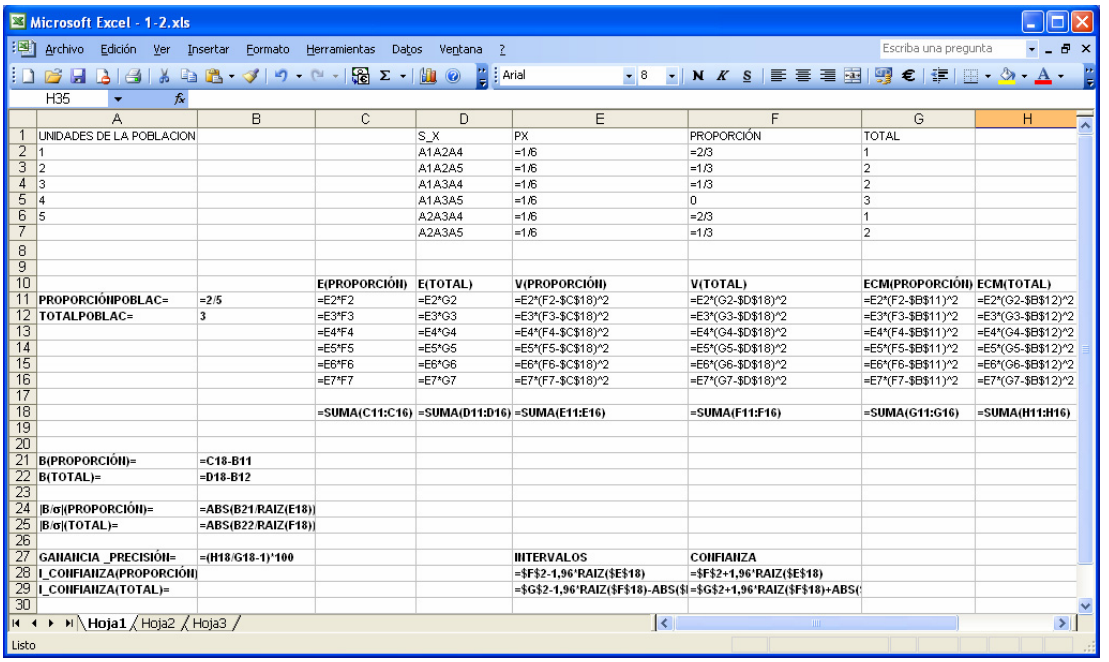
$$\left[\hat{T}_1 - \frac{\sigma(\hat{T}_1)}{\sqrt{\alpha}}, \hat{T}_1 + \frac{\sigma(\hat{T}_1)}{\sqrt{\alpha}}\right] = \left[2/3 - \frac{0{,}229}{\sqrt{0{,}05}}, 2/3 + \frac{0{,}229}{\sqrt{0{,}05}}\right] = [0.357, 1.69]$$

Se observa que el intervalo de confianza para $\hat{T}_1$ sin existir normalidad es más ancho, es decir, es menos preciso.

Para hallar un intervalo de confianza para $\hat{T}_2$ (que es sesgado) basado en la primera muestra, realizamos los siguientes cálculos:

$$[\hat{T}_2 - \lambda_\alpha \sigma(\hat{T}_2) - |B(\hat{T}_2)|, \hat{T}_2 + \lambda_\alpha \sigma(\hat{T}_2) + |B(\hat{T}_2)|] = [1 - 1.96*0.687 + 0.16, 1 + 1.96*0.687 + 0.16] = [-1.513, 3.513]$$

Se observa que el intervalo de confianza del estimador menos preciso es más ancho. Los cálculos pueden automatizarse con Excel como sigue:



Microsoft Excel - 1-2.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

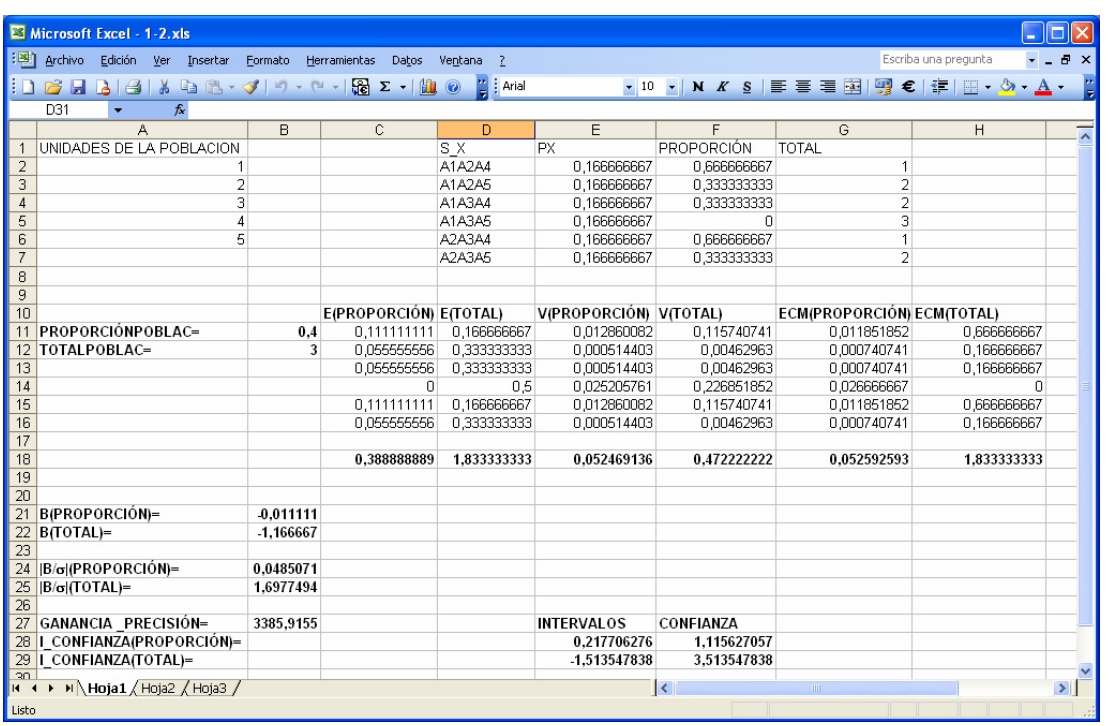
Arial 8

H35

	A	B	C	D	E	F	G	H
1	UNIDADES DE LA POBLACION			S_X	PX	PROPORCIÓN	TOTAL	
2	1			A1A2A4	=1/6	=2/3	1	
3	2			A1A2A5	=1/6	=1/3	2	
4	3			A1A3A4	=1/6	=1/3	2	
5	4			A1A3A5	=1/6	0	3	
6	5			A2A3A4	=1/6	=2/3	1	
7				A2A3A5	=1/6	=1/3	2	
8								
9								
10			E(PROPORCIÓN)	E(TOTAL)	V(PROPORCIÓN)	V(TOTAL)	ECM(PROPORCIÓN)	ECM(TOTAL)
11	PROPORCIÓNPOBLAC=	=2/5	=E2*F2	=E2*G2	=E2*(F2-$C$18)^2	=E2*(G2-$D$18)^2	=E2*(F2-$B$11)^2	=E2*(G2-$B$12)^2
12	TOTALPOBLAC=	3	=E3*F3	=E3*G3	=E3*(F3-$C$18)^2	=E3*(G3-$D$18)^2	=E3*(F3-$B$11)^2	=E3*(G3-$B$12)^2
13			=E4*F4	=E4*G4	=E4*(F4-$C$18)^2	=E4*(G4-$D$18)^2	=E4*(F4-$B$11)^2	=E4*(G4-$B$12)^2
14			=E5*F5	=E5*G5	=E5*(F5-$C$18)^2	=E5*(G5-$D$18)^2	=E5*(F5-$B$11)^2	=E5*(G5-$B$12)^2
15			=E6*F6	=E6*G6	=E6*(F6-$C$18)^2	=E6*(G6-$D$18)^2	=E6*(F6-$B$11)^2	=E6*(G6-$B$12)^2
16			=E7*F7	=E7*G7	=E7*(F7-$C$18)^2	=E7*(G7-$D$18)^2	=E7*(F7-$B$11)^2	=E7*(G7-$B$12)^2
17								
18			=SUMA(C11:C16)	=SUMA(D11:D16)	=SUMA(E11:E16)	=SUMA(F11:F16)	=SUMA(G11:G16)	=SUMA(H11:H16)
19								
20								
21	B(PROPORCIÓN)=	=C18-B11						
22	B(TOTAL)=	=D18-B12						
23								
24	|B/σ|(PROPORCIÓN)=	=ABS(B21/RAIZ(E18))						
25	|B/σ|(TOTAL)=	=ABS(B22/RAIZ(F18))						
26								
27	GANANCIA _PRECISIÓN=	=(H18/G18-1)*100			INTERVALOS	CONFIANZA		
28	I_CONFIANZA(PROPORCIÓN)				=$F$2-1,96*RAIZ($E$18)	=$F$2+1,96*RAIZ($E$18)		
29	I_CONFIANZA(TOTAL)=				=$G$2-1,96*RAIZ($F$18)-ABS($	=$G$2+1,96*RAIZ($F$18)+ABS(		
30								

Hoja1 Hoja2 Hoja3

Listo



Microsoft Excel - 1-2.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

Arial 10

D31

	A	B	C	D	E	F	G	H
1	UNIDADES DE LA POBLACION			S_X	PX	PROPORCIÓN	TOTAL	
2	1			A1A2A4	0,166666667	0,666666667	1	
3	2			A1A2A5	0,166666667	0,333333333	2	
4	3			A1A3A4	0,166666667	0,333333333	2	
5	4			A1A3A5	0,166666667	0	3	
6	5			A2A3A4	0,166666667	0,666666667	1	
7				A2A3A5	0,166666667	0,333333333	2	
8								
9								
10			E(PROPORCIÓN)	E(TOTAL)	V(PROPORCIÓN)	V(TOTAL)	ECM(PROPORCIÓN)	ECM(TOTAL)
11	PROPORCIÓNPOBLAC=	0,4	0,111111111	0,166666667	0,012860082	0,115740741	0,011851852	0,666666667
12	TOTALPOBLAC=	3	0,055555556	0,333333333	0,000514403	0,00462963	0,000740741	0,166666667
13			0,055555556	0,333333333	0,000514403	0,00462963	0,000740741	0,166666667
14			0	0,5	0,025205761	0,226851852	0,026666667	0
15			0,111111111	0,166666667	0,012860082	0,115740741	0,011851852	0,666666667
16			0,055555556	0,333333333	0,000514403	0,00462963	0,000740741	0,166666667
17								
18			**0,388888889**	**1,833333333**	**0,052469136**	**0,472222222**	**0,052592593**	**1,833333333**
19								
20								
21	B(PROPORCIÓN)=	-0,011111						
22	B(TOTAL)=	-1,166667						
23								
24	|B/σ|(PROPORCIÓN)=	0,0485071						
25	|B/σ|(TOTAL)=	1,6977494						
26								
27	GANANCIA _PRECISIÓN=	3385,9155			INTERVALOS	CONFIANZA		
28	I_CONFIANZA(PROPORCIÓN)=				0,217706276	1,115627057		
29	I_CONFIANZA(TOTAL)=				-1,513547838	3,513547838		
30								

Hoja1 Hoja2 Hoja3

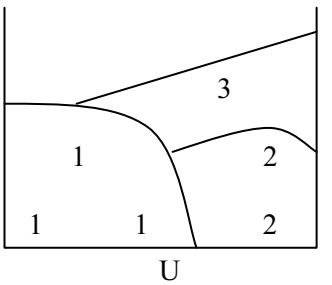
Listo

1.3. En una población de 3 unidades numeradas $\{U_1, U_2, U_3\}$ se extraen muestras de tamaño 2 mediante el siguiente método de muestreo: Se extraen al azar 2 bolas de una urna que contiene 6 bolas (tres con el número 1, dos con el número 2 y una con el número 3), y se extraen de la población las dos unidades que tengan los mismos números que las dos bolas extraídas. Se pide:

1) Considerando la extracción de las bolas en la urna con reposición y el estimador por analogía T= Número de unidades distintas en las muestras, hallar su distribución en el muestreo analizando su precisión. Obtener una estimación puntual del número de unidades distintas en la población y otra por intervalos al 99,8% de confianza ($F^{-1}(0{,}999) = 3$) basándose en la muestra de mayor probabilidad.

2) Contestar a las preguntas del apartado anterior suponiendo que la extracción de las bolas en la urna sin reposición. Comparar las estimaciones en los dos casos comentando los resultados.

Para hallar el espacio muestral asociado a este procedimiento de muestreo ***sin reposición*** consideramos la urna U con 6 bolas (tres con el número 1, dos con el número 2 y una con el número 3).



Como en la urna U seleccionamos dos bolas sin reposición, las posibilidades son (1,1), (1,2), (1,3), (2,2) y (2,3).

Las probabilidades de las muestras se calculan como se indica a continuación:

$$P(1,1) = P_1(1) + P_2(1/1) = \frac{3}{6}\cdot\frac{2}{5} = \frac{1}{5}$$

$$P(1,2) = P\{1,2\} + P\{2,1\} = P_1(1)P_2(2/1) + P_1(2)P_2(1/2) = \frac{3}{6}\cdot\frac{2}{5} + \frac{2}{6}\cdot\frac{3}{5} = \frac{2}{5}$$

$$P(1,3) = P\{1,3\} + P\{3,1\} = P_1(1)P_2(3/1) + P_1(3)P_2(1/3) = \frac{3}{6}\cdot\frac{1}{5} + \frac{1}{6}\cdot\frac{3}{5} = \frac{1}{5}$$

$$P(2,2) = P_1(2)\cdot P_2(2/2) = \frac{2}{6}\cdot\frac{1}{5} = \frac{1}{15}$$

$$P(2,3) = P\{2,3\} + P\{3,2\} = P_1(2)P_2(3/2) + P_1(3)P_2(2/3) = \frac{2}{6}\cdot\frac{1}{5} + \frac{1}{6}\cdot\frac{2}{5} = \frac{2}{15}$$

Los índices 1 y 2 de las probabilidades indican primera y segunda extracción, respectivamente. Las barras inclinadas indican condicionada a que se haya obtenido en la primera extracción el número que aparece en el denominador.

Ya podemos formar la tabla con las muestras del espacio muestral *S_X*, sus probabilidades P_i y los valores del estimador *T* del problema sobre las mismas, datos que nos van a permitir el cálculo de la distribución en el muestreo del estimador. En el siguiente cuadro se especifican las muestras, sus probabilidades y los valores del estimador para cada muestra.

Muestras (*sin reposicón*)	*S _ X*	*Pi*	*T*
1	(1,1)	1/5	1
2	(1,2)	2/5	2
3	(1,3)	1/5	2
4	(2,2)	1/15	1
5	(2,3)	2/15	2

La distribución de probabilidad del estimador en el muestreo se calcularán mediante la expresión ya conocida $P^T(\hat{\theta}(X_1, ...,X_n) = t) = \sum_{\{S_i / \hat{\theta}(S_i(X))=t\}} P(S_i)$, de la siguiente forma:

$$T\begin{cases} P^T(T=1) = \dfrac{1}{5} + \dfrac{1}{15} = \dfrac{4}{15} \\ P^T(T=2) = \dfrac{2}{5} + \dfrac{1}{5} + \dfrac{2}{15} = \dfrac{11}{15} \end{cases}$$

Una vez conocida la distribución de probabilidad en el muestreo del estimador analizaremos si es insesgado o no. Para ello observamos que el valor del número de unidades distintas en la población es $\theta = 3$, que es el parámetro que estamos estimando con el estimador T.

Ahora, para comprobar la insesgadez, hallamos la esperanza matemática del estimador tal y como se indica a continuación:

$$E(T) = 1 \cdot \frac{4}{15} + 2 \cdot \frac{11}{15} = 26/15 = 1{,}7333333333 \neq 3 = \theta$$

El estimador *T* es sesgado con sesgo $B(T)=E(T) - \theta = 26/15\text{-}3 = -19/15 = -1{,}26666$. La varianza del estimador es la siguiente:

$$V(T) = (1 - 1{,}733)^2 \cdot \frac{4}{15} + (2 - 1{,}733)^2 \cdot \frac{11}{15} = 0{,}1955$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(T) = \sqrt{0{,}1955} = 0{,}442$$

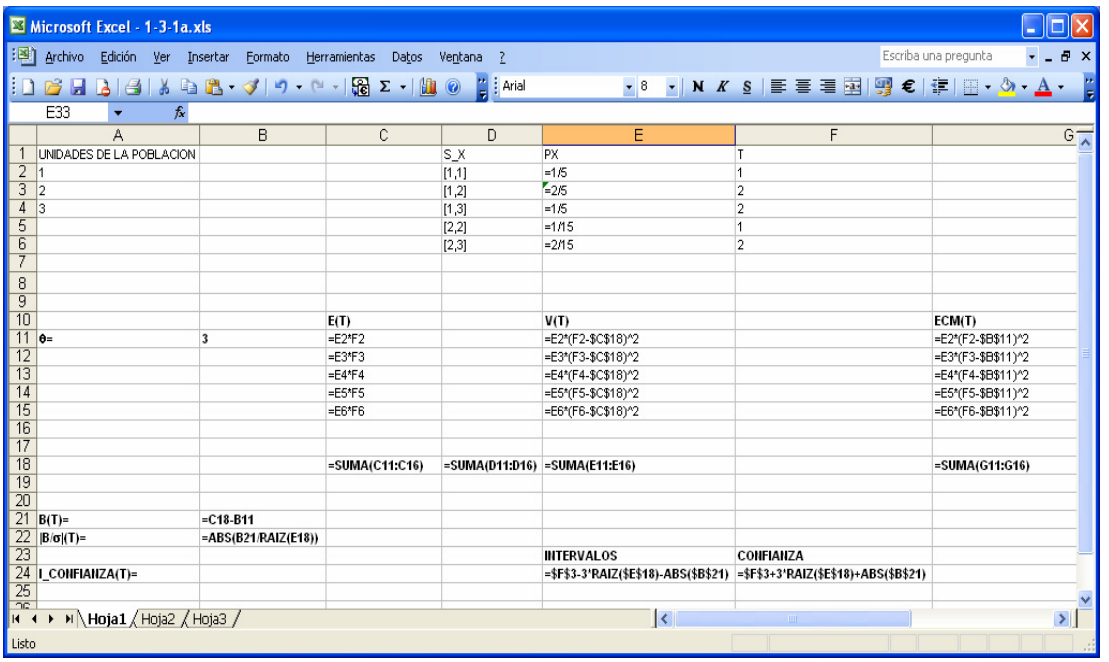
Como $|B(T)/\sigma(T)| = 1{,}266/0{,}442 = 2{,}864 > 1/10$, el sesgo del estimador *T* no es despreciable, por lo que calcularemos su precisión mediante el error cuadrático medio. Tenemos:

$$ECM(T) = (1 - 3)^2 \cdot \frac{4}{15} + (2 - 3)^2 \cdot \frac{11}{15} = 1{,}8$$

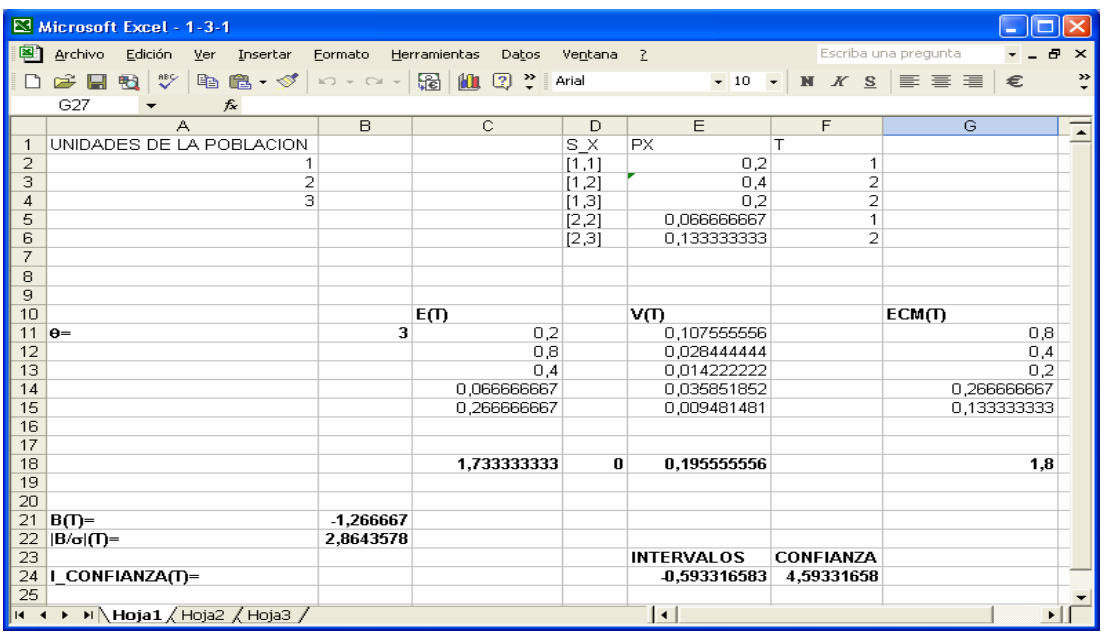
Para hallar un intervalo de confianza para T (que es sesgado) basado en la segunda muestra (que es la de mayor probabilidad), realizamos los siguientes cálculos:

$$[T-\lambda_\alpha\sigma(T)-|B(T)|, T+\lambda_\alpha\sigma(T)+|B(T)|] = [2-3*0{,}442-1{,}26+3*0{,}442+1{,}26] = [-0{,}593, 4{,}593]$$

Los cálculos pueden automatizarse con Excel como sigue:



	A	B	C	D	E	F	G
1	UNIDADES DE LA POBLACION			S_X	PX	T	
2	1			[1,1]	=1/5	1	
3	2			[1,2]	=2/5	2	
4	3			[1,3]	=1/5	2	
5				[2,2]	=1/15	1	
6				[2,3]	=2/15	2	
7							
8							
9							
10			E(T)		V(T)		ECM(T)
11	θ=	3	=E2*F2		=E2*(F2-$C$18)^2		=E2*(F2-$B$11)^2
12			=E3*F3		=E3*(F3-$C$18)^2		=E3*(F3-$B$11)^2
13			=E4*F4		=E4*(F4-$C$18)^2		=E4*(F4-$B$11)^2
14			=E5*F5		=E5*(F5-$C$18)^2		=E5*(F5-$B$11)^2
15			=E6*F6		=E6*(F6-$C$18)^2		=E6*(F6-$B$11)^2
16							
17							
18			=SUMA(C11:C16)	=SUMA(D11:D16)	=SUMA(E11:E16)		=SUMA(G11:G16)
19							
20							
21	B(T)=	=C18-B11					
22	|B/σ|(T)=	=ABS(B21/RAIZ(E18))					
23					INTERVALOS	CONFIANZA	
24	I_CONFIANZA(T)=				=$F$3-3*RAIZ($E$18)-ABS($B$21)	=$F$3+3*RAIZ($E$18)+ABS($B$21)	
25							



	A	B	C	D	E	F	G
1	UNIDADES DE LA POBLACION			S_X	PX	T	
2	1			[1,1]	0,2	1	
3	2			[1,2]	0,4	2	
4	3			[1,3]	0,2	2	
5				[2,2]	0,066666667	1	
6				[2,3]	0,133333333	2	
7							
8							
9							
10			E(T)		V(T)		ECM(T)
11	θ=	3	0,2		0,107555556		0,8
12			0,8		0,028444444		0,4
13			0,4		0,014222222		0,2
14			0,066666667		0,035851852		0,266666667
15			0,266666667		0,009481481		0,133333333
16							
17							
18			1,733333333	0	0,195555556		1,8
19							
20							
21	B(T)=	-1,266667					
22	|B/σ|(T)=	2,8643578					
23					INTERVALOS	CONFIANZA	
24	I_CONFIANZA(T)=				-0,593316583	4,59331658	
25							

Cuando en la urna U seleccionamos dos bolas ***con reposición***, las posibilidades son (1,1), (1,2), (1,3), (2,2), (2,3) y (3,3).

Las probabilidades de las muestras se calculan como se indica a continuación:

$$P(1,1) = P(1)\cdot P(1) = \frac{3}{6}\cdot\frac{3}{6} = \frac{1}{4}$$

$$P(1,2) = 2P(1)\cdot P(2) = 2\,\frac{3}{6}\cdot\frac{2}{5} = \frac{1}{3}$$

$$P(1,3) = 2P(1)\cdot P(3) = 2\,\frac{3}{6}\cdot\frac{1}{6} = \frac{1}{6}$$

$$P(2,2) = P(2)\cdot P(2) = \frac{2}{6}\cdot\frac{2}{6} = \frac{1}{9}$$

$$P(2,3) = 2P(2)\cdot P(3) = 2\,\frac{2}{6}\cdot\frac{1}{6} = \frac{1}{9}$$

$$P(3,3) = P(3)\cdot P(3) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$$

Ya podemos formar la tabla con las muestras del espacio muestral *S_X*, sus probabilidades P_i y los valores del estimador *T* del problema sobre las mismas, datos que nos van a permitir el cálculo de la distribución en el muestreo del estimador. En el siguiente cuadro se especifican las muestras, sus probabilidades y los valores del estimador para cada muestra.

Muestras (*con reposición*)	*S _ X*	*Pi*	*T*
1	(1,1)	1 / 4	1
2	(1,2)	1 / 3	2
3	(1,3)	1 / 6	2
4	(2,2)	1 / 9	1
5	(2,3)	1 / 9	2
6	(3,3)	1 / 36	1

La distribución de probabilidad del estimador en el muestreo se calcularán mediante la expresión ya conocida $P^T(\hat{\theta}(X_1, ...,X_n) = t) = \sum_{\{S_i / \hat{\theta}(S_i(X))=t\}} P(S_i)$, de la siguiente forma:

$$T\begin{cases} P^T(T=1) = \frac{1}{4}+\frac{1}{9}+\frac{1}{36} = \frac{7}{18} \\ P^T(T=2) = \frac{1}{3}+\frac{1}{6}+\frac{1}{9} = \frac{11}{18} \end{cases}$$

Una vez conocida la distribución de probabilidad en el muestreo del estimador analizaremos si es insesgado o no. Para ello observamos que el valor del número de unidades distintas en la población es $\theta = 3$, que es el parámetro que estamos estimando con el estimador *T*.

Ahora, para comprobar la insesgadez, hallamos la esperanza matemática del estimador tal y como se indica a continuación:

$$E(T) = 1 \cdot \frac{7}{18} + 2 \cdot \frac{11}{18} = 29/18 = 1{,}611111 \neq 3 = \theta$$

El estimador T es sesgado con sesgo B(T)=E(T)-θ = 29/18-3 = -25/18=-1,3888. La varianza del estimador es la siguiente:

$$V(T) = (1 - 1{,}6111)^2 \cdot \frac{7}{18} + (2 - 1{,}6111)^2 \cdot \frac{11}{18} = 0{,}237$$

Con lo que las desviaciones típicas valdrá:

$$\sigma(T) = \sqrt{0{,}237} = 0{,}486$$

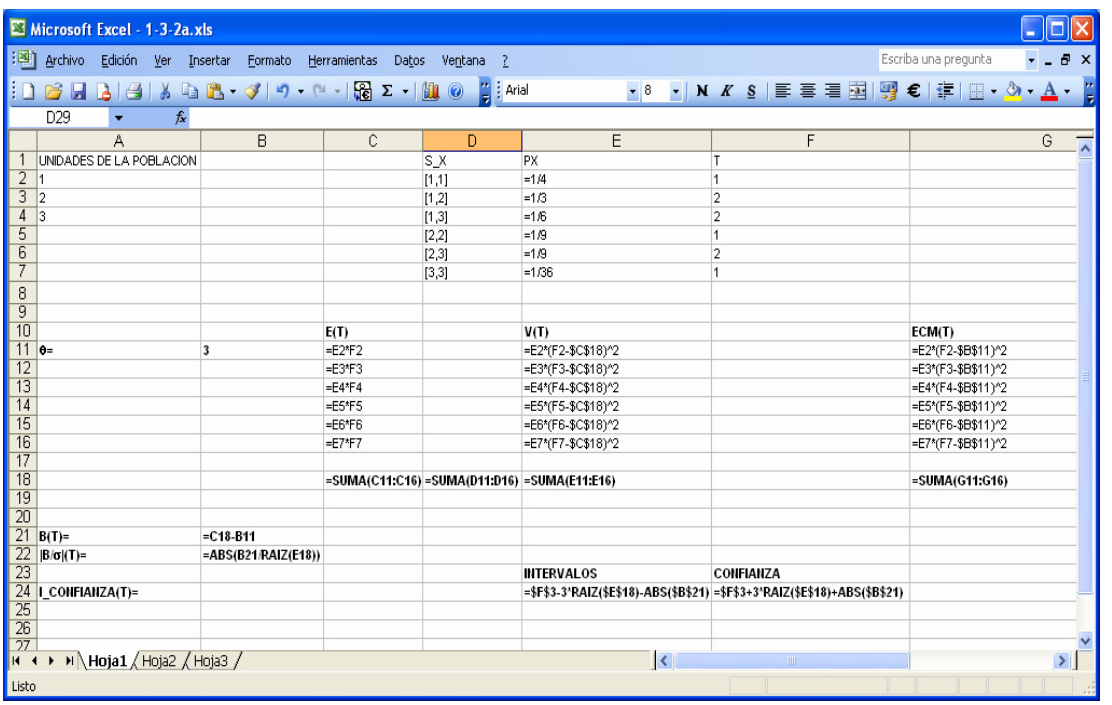
Como |B(T)/σ(T)| = 1,388/0,486 = 2,85 > 1/10, el sesgo del estimador T no es despreciable, por lo que calcularemos su precisión mediante el error cuadrático medio. Tenemos:

$$ECM(T) = (1 - 3)^2 \cdot \frac{7}{18} + (2 - 3)^2 \cdot \frac{11}{18} = 2{,}1666$$

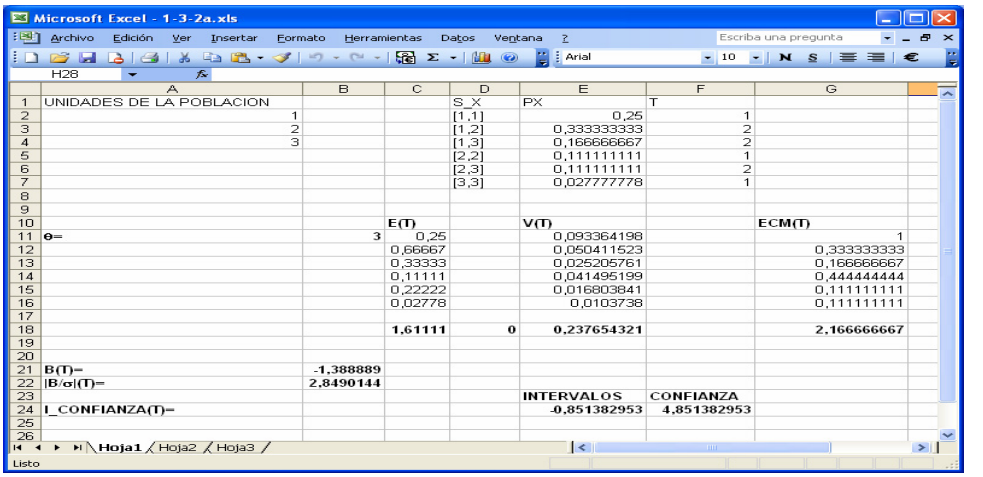
Para hallar un intervalo de confianza para T (que es sesgado) basado en la segunda muestra (que es la de mayor probabilidad), realizamos los siguientes cálculos:

$$[T - \lambda_\alpha \sigma(T) - |B(T)|, T + \lambda_\alpha \sigma(T) + |B(T)|] = [2 - 3*0{,}486 - 1{,}38 + 3*0{,}486 + 1{,}38] = [-0{,}851, 4{,}851]$$

Los cálculos pueden automatizarse con Excel como sigue:



	A	B	C	D	E	F	G
1	UNIDADES DE LA POBLACION			S_X	PX	T	
2	1			[1,1]	=1/4	1	
3	2			[1,2]	=1/3	2	
4	3			[1,3]	=1/6	2	
5				[2,2]	=1/9	1	
6				[2,3]	=1/9	2	
7				[3,3]	=1/36	1	
8							
9							
10			E(T)		V(T)		ECM(T)
11	θ=	3	=E2*F2		=E2*(F2-$C$18)^2		=E2*(F2-$B$11)^2
12			=E3*F3		=E3*(F3-$C$18)^2		=E3*(F3-$B$11)^2
13			=E4*F4		=E4*(F4-$C$18)^2		=E4*(F4-$B$11)^2
14			=E5*F5		=E5*(F5-$C$18)^2		=E5*(F5-$B$11)^2
15			=E6*F6		=E6*(F6-$C$18)^2		=E6*(F6-$B$11)^2
16			=E7*F7		=E7*(F7-$C$18)^2		=E7*(F7-$B$11)^2
17							
18			=SUMA(C11:C16)	=SUMA(D11:D16)	=SUMA(E11:E16)		=SUMA(G11:G16)
19							
20							
21	B(T)=	=C18-B11					
22	|B/σ|(T)=	=ABS(B21/RAIZ(E18))					
23					INTERVALOS	CONFIANZA	
24	I_CONFIANZA(T)=				=$F$3-3*RAIZ($E$18)-ABS($B$21)	=$F$3+3*RAIZ($E$18)+ABS($B$21)	
25							
26							
27							



Microsoft Excel - 1-3-2a.xls

	A	B	C	D	E	F	G
1	UNIDADES DE LA POBLACION			S_X	PX	T	
2	1			[1,1]	0,25	1	
3	2			[1,2]	0,333333333	2	
4	3			[1,3]	0,166666667	2	
5				[2,2]	0,111111111	1	
6				[2,3]	0,111111111	2	
7				[3,3]	0,027777778	1	
8							
9							
10			E(T)		V(T)		ECM(T)
11	θ=	3	0,25		0,093364198		1
12			0,66667		0,050411523		0,333333333
13			0,33333		0,025205761		0,166666667
14			0,11111		0,041496199		0,444444444
15			0,22222		0,016803841		0,111111111
16			0,02778		0,0103738		0,111111111
17							
18			1,61111	0	0,237654321		2,166666667
19							
20							
21	B(T)=	-1,388889					
22	|B/σ|(T)=	2,8490144					
23					INTERVALOS	CONFIANZA	
24	I_CONFIANZA(T)=				-0,851382953	4,851382953	
25							
26							

Hoja1 / Hoja2 / Hoja3 /

Listo

Para comparar las estimaciones con y sin reposición observamos los errores cuadráticos medios, resultando que el método sin reposición tiene menor error cuadrático medio, lo que indica que es mejor método de estimación.

La ganancia en precisión por trabajar sin reposición en vez de con reposición se cuantifica como sigue:

$$GP = (EMC_{CR}(T)/EMC_{SR}(T) - 1)*100 = (2,1666/1,8-1)*100=20,37\%$$

Se ve que la precisión mejora un 20,37% en caso de usa selección sin reposición. Además, también se observa que el intervalo de confianza del estimador menos preciso (con reposición) es más ancho.

1.4. Con la finalidad de ensayar el análisis de la divisibilidad en una población numérica, consideramos una población virtual finita con 6 elementos $U = \{12, 13, 17, 23, 6, 1\}$. Mediante un método de muestreo aleatorio con probabilidades iguales y sin reposición se extraen muestras de tamaño 2 sin tener en cuenta el orden de colocación de sus elementos.

1) ¿Cuántos elementos tiene el espacio muestral? Especificar dicho espacio muestral y las probabilidades asociadas a las muestras.

2) A partir de las muestras del espacio muestral se trata de estimar el parámetro poblacional PROPORCIÓN DE NÚMEROS PRIMOS mediante el estimador por analogía y el parámetro poblacional TOTAL DE NÚMEROS PRIMOS mediante el estimador de expansión de la proporción por el tamaño poblacional (producto del estimador de la proporción por el tamaño poblacional). Hallar la distribución en el muestreo de dichos estimadores. ¿Qué estimador es mejor? Comparar el estimador de expansión del total con el estimador por analogía.

3) Hallar intervalos de confianza al 99% ($\alpha = 0.01$) para el total y la proporción de números primos en la población, basados en las muestras cuyos dos elementos son números no primos. Tenemos como dato conocido que $F^{-1}(0.995)= 2.57$, siendo F la función de distribución de la normal (0,1). Comentar los resultados.

Como se trata de muestreo aleatorio sin reposición en el que el orden de colocación de los elementos en las muestras de tamaño 2 no interviene, el número de muestras posibles será:

$$\binom{6}{2}=15$$

Por otra parte, en este problema estamos considerando la clase A de los números primos, con lo que asociaremos a los U_i los A_i que valen cero cuando U_i no es primo y valen uno cuando U_i es primo. Luego sobre el conjunto $U_i \rightarrow \{2, 13, 17, 23, 6, 1\}$ se mide la variable A y se obtiene el conjunto $A_i \rightarrow \{0, 1, 1, 1, 0, 1\}$. Al tratarse de muestreo aleatorio sin reposición y probabilidades iguales, las probabilidades iniciales de selección de los elementos de la población para la muestra valdrán $P(u_i) = 1/6$, $i = 1, ..., 6$ y la probabilidad de cualquier muestra puede hallarse mediante la expresión:

$$P_X = \mathrm{P}(u_i,u_j) = \mathrm{P}(u_i)\mathrm{P}(u_j/u_i) + \mathrm{P}(u_j)\mathrm{P}(u_i/u_j) = \mathrm{P}(u_i)\mathrm{P}(u_j)/(1-\mathrm{P}(u_i)) + \mathrm{P}(u_j)\mathrm{P}(u_i)/(1-\mathrm{P}(u_j))) =$$
$$(1/6^2)/(1-1/6) + (1/6^2)/(1-1/6) = 2(1/6^2)/(1-1/6) = 1/15$$

Se observa que las probabilidades de las muestras serán todas iguales a 1/15. Luego estamos ante un método de selección con probabilidades iguales y muestras equiprobables.

A continuación se presenta la tabla que contiene el espacio muestral, las probabilidades de las muestras y la distribución de los estimadores.

S1_X	S2_X	P_X	PROPORCIÓN ($\hat{P}$)	TOTAL ($\hat{A}=6\hat{P}$) EXPANSIÓN	TOTAL ($\hat{T}=2\hat{P}$) MUESTRAL
0	1	1/15	0,5	3	1
0	1	1/15	0,5	3	1
0	1	1/15	0,5	3	1
0	0	1/15	0	0	0
0	1	1/15	0,5	3	1
1	1	1/15	1	6	2
1	1	1/15	1	6	2
1	0	1/15	0,5	3	1
1	1	1/15	1	6	2
1	1	1/15	1	6	2
1	0	1/15	0,5	3	1
1	1	1/15	1	6	2
1	0	1/15	0,5	3	1
1	1	1/15	1	6	2
0	1	1/15	0,5	3	1

Las distribuciones de probabilidad de los dos estimadores se calcularán mediante la expresión ya conocida $P^T(\hat{\theta}(X_1, ..., X_n) = t) = \sum_{\{S_i / \hat{\theta}(S_i(X))=t\}} P(S_i)$, de la siguiente forma:

$$\hat{P}\begin{cases} P^T(\hat{P}=1) = 6\cdot\frac{1}{15} = \frac{2}{5} \\ P^T(\hat{P}=1/2) = 8\cdot\frac{1}{15} = \frac{8}{15} \\ P^T(\hat{P}=0) = \frac{1}{15} \end{cases} \quad \hat{A}\begin{cases} P^T(\hat{A}=6) = 6\cdot\frac{1}{15} = \frac{2}{5} \\ P^T(\hat{A}=3) = 8\cdot\frac{1}{15} = \frac{8}{15} \\ P^T(\hat{A}=0) = \frac{1}{15} \end{cases} \quad \hat{T}\begin{cases} P^T(\hat{T}=2) = 6\cdot\frac{1}{15} = \frac{2}{5} \\ P^T(\hat{T}=1) = 8\cdot\frac{1}{15} = \frac{8}{15} \\ P^T(\hat{T}=0) = \frac{1}{15} \end{cases}$$

Una vez conocida la distribución de probabilidad en el muestreo de los estimadores analizaremos si son insesgados o no. Para ello calculamos en primer lugar los valores de la proporción de números primos de la población $\theta_1 = 2/3$ y del total de números primos de la población $\theta_2 = 4$.

Ahora, para comprobar la insesgadez, hallamos la esperanza matemática de los estimadores tal y como se indica a continuación:

$$E(\hat{P}) = 1\cdot\frac{2}{5} + \frac{1}{2}\cdot\frac{8}{15} + 0\cdot\frac{1}{15} = 2/3 = 0{,}6666 = \theta_1$$

$$E(\hat{A}) = 6\cdot\frac{2}{5} + 3\cdot\frac{8}{15} + 0\cdot\frac{1}{15} = 6E(\hat{P}) = 4 = \theta_2$$

$$E(\hat{T}) = 2\cdot\frac{2}{5} + 1\cdot\frac{8}{15} + 0\cdot\frac{1}{15} = 2E(\hat{P}) = 4/3 = 1{,}33333 \neq 4 = \theta_2$$

Se observa que $\hat{P}$ es insesgado para θ_1 y $\hat{A}$ e insesgado para θ_2. El estimador $\hat{T}$ es sesgado para θ_2 con sesgo $B(\hat{T}) = E(\hat{T}) - \theta_2 = 4/3 - 4 = -8/3 = -2{,}66$. Para calcular las varianzas de los estimadores se tiene en cuenta que $\hat{A} = 6\hat{P}$ y que $\hat{T} = 2\hat{P}$.

$$V(\hat{P}) = (1-0{,}666)^2\cdot\frac{2}{5} + (\frac{1}{2}-0{,}666)^2\cdot\frac{8}{15} + (0-0{,}66)^2\cdot\frac{1}{15} = 0{,}088888$$

$$V(\hat{A}) = V(6\hat{P}) = 36V(\hat{P}) = 3{,}2$$

$$V(\hat{T}) = V(2\hat{P}) = 4V(\hat{P}) = 0{,}35555$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(\hat{P}) = \sqrt{0{,}088888} = 0{,}298, \quad \sigma(\hat{A}) = \sqrt{3{,}2} = 1{,}7888 y \quad \sigma(\hat{T}) = \sqrt{0{,}35555} = 0{,}596$$

Como los estimadores $\hat{P}$ y $\hat{A}$ son insesgados, su varianza coincide con su error cuadrático medio, por lo que su precisión se mide a través de la varianza. De esta forma, el estimador $\hat{P}$ para estimar θ_1 es más preciso que el estimador $\hat{A}$ para estimar θ_2 por tener menor varianza.

Como $|B(\hat{T})/\sigma(\hat{T})| = 4{,}46 > 1/10$, el sesgo del estimador $\hat{T}$ no es despreciable y al compararlo con $\hat{A}$ tenemos un estimador sesgado y el otro insesgado. La comparación debe hacerse a través de los errores cuadráticos medios. Tenemos:

$$ECM(\hat{T}) = (1-4/3)^2\cdot\frac{2}{5} + (\frac{1}{2}-4/3)^2\cdot\frac{8}{15} + (0-4/3)^2\cdot\frac{1}{15} = 0{,}53333 > ECM(\hat{A}) = V(\hat{A}) = 3{,}2$$

Se observa que el error cuadrático medio de $\hat{T}$ es mayor que la varianza de $\hat{A}$, luego $\hat{A}$ es más preciso que $\hat{T}$ para estimar θ_2. Por lo tanto, el estimador de expansión del total es más preciso que el estimador por analogía.

Para el cálculo de las estimaciones por intervalos (intervalos de confianza de los estimadores) es útil poder suponer que A se distribuye normalmente. Como el coeficiente de asimetría de A vale −0,96 y el de curtosis −1,87, puede ser lógico suponer la normalidad, ya que ambos coeficientes se encuentran en el intervalo [−2,2]. Sin embargo, como el extremo inferior está muy cerca de −2, para aceptar esta suposición será necesario realizar un contraste formal de normalidad. Por lo tanto, hallamos los intervalos de confianza bajo las dos hipótesis (normalidad y no normalidad en la población).

Los coeficientes de asimetría g_1 y curtosis g_2 de A se calculan como sigue:

$$g_1 = \frac{m3}{\sigma^3} = \frac{\frac{1}{6}\left[2(0-2/3)^3 + 4(1-2/3)^3\right]}{\left(\sqrt{\frac{1}{6}\left[2(0-2/3)^2 + 4(1-2/3)^2\right]}\right)^3} = 0{,}968$$

$$g_2 = \frac{m4}{\sigma^4} - 3 = \frac{\frac{1}{6}\left[2(0-2/3)^4 + 4(1-2/3)^4\right]}{\left(\sqrt{\frac{1}{6}\left[2(0-2/3)^2 + 4(1-2/3)^2\right]}\right)^4} - 3 = -1{,}875$$

Supuesta la no normalidad de A, para hallar un intervalo de confianza para la proporción $\hat{P}$ al 99%, basado en la única muestra (0,0) correspondiente al único par de elementos ambos no primos (12,6), utilizamos el intervalo:

$$\left[\hat{P} - \frac{\sigma(\hat{P})}{\sqrt{\alpha}},\ \hat{P} + \frac{\sigma(\hat{P})}{\sqrt{\alpha}}\right] = \left[0 - \frac{0{,}298}{\sqrt{0{,}01}},\ 0 + \frac{0{,}298}{\sqrt{0{,}01}}\right] = \left[-2.98,\ 2.98\right]$$

Si se hubiera supuesto normalidad el intervalo de confianza para $\hat{P}$ al 99% sería:

$$[\hat{P} - \lambda_\alpha \sigma(\hat{P}),\ \hat{P} + \lambda_\alpha \sigma(\hat{P})] = [0 - 2{,}57 \cdot 0{,}298,\ 0 + 2{,}57 \cdot 0{,}298] = \left[-0.766,\ 0.766\right]$$

Se observa que el intervalo de confianza en presencia de normalidad es más estrecho (más preciso) que sin normalidad.

Dada la no normalidad de A, para hallar un intervalo de confianza para el total de clase $\hat{A}$ al 99%, basado en la única muestra (0,0) correspondiente al único par de elementos ambos no primos (12,6), utilizamos el intervalo:

$$\left[\hat{A} - \frac{\sigma(\hat{A})}{\sqrt{\alpha}},\ \hat{A} + \frac{\sigma(\hat{A})}{\sqrt{\alpha}}\right] = \left[0 - \frac{1{,}7888}{\sqrt{0{,}01}},\ 0 + \frac{1{,}7888}{\sqrt{0{,}01}}\right] = \left[-17.8,\ 17.8\right]$$

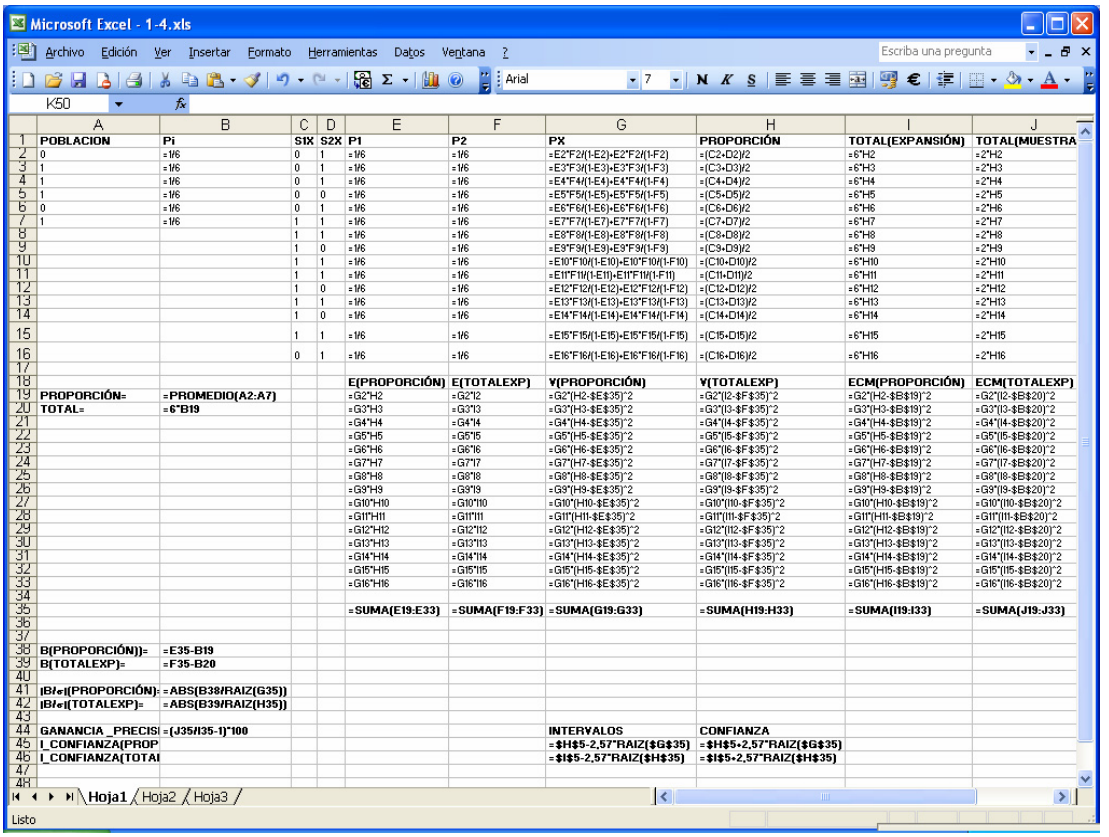
Si se hubiera supuesto normalidad el intervalo de confianza para $\hat{A}$ al 99% sería:

$$[\hat{A} - \lambda_\alpha \sigma(\hat{A}),\ \hat{A} + \lambda_\alpha \sigma(\hat{A})] = [0 - 2{,}57 \cdot 1{,}7888,\ 0 + 2{,}57 \cdot 1{,}7888] = \left[-4.59,\ 4{,}59\right]$$

Se observa que el intervalo de confianza en presencia de normalidad es más estrecho (más preciso) que sin normalidad.

Además, se observa que los intervalos de confianza para $\hat{P}$ son más estrechos que los correspondientes intervalos de confianza para $\hat{A}$, lo que concuerdo con la superior precisión del estimador $\hat{P}$.

Los cálculos pueden automatizarse con Excel como sigue:



Microsoft Excel - 1-4.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

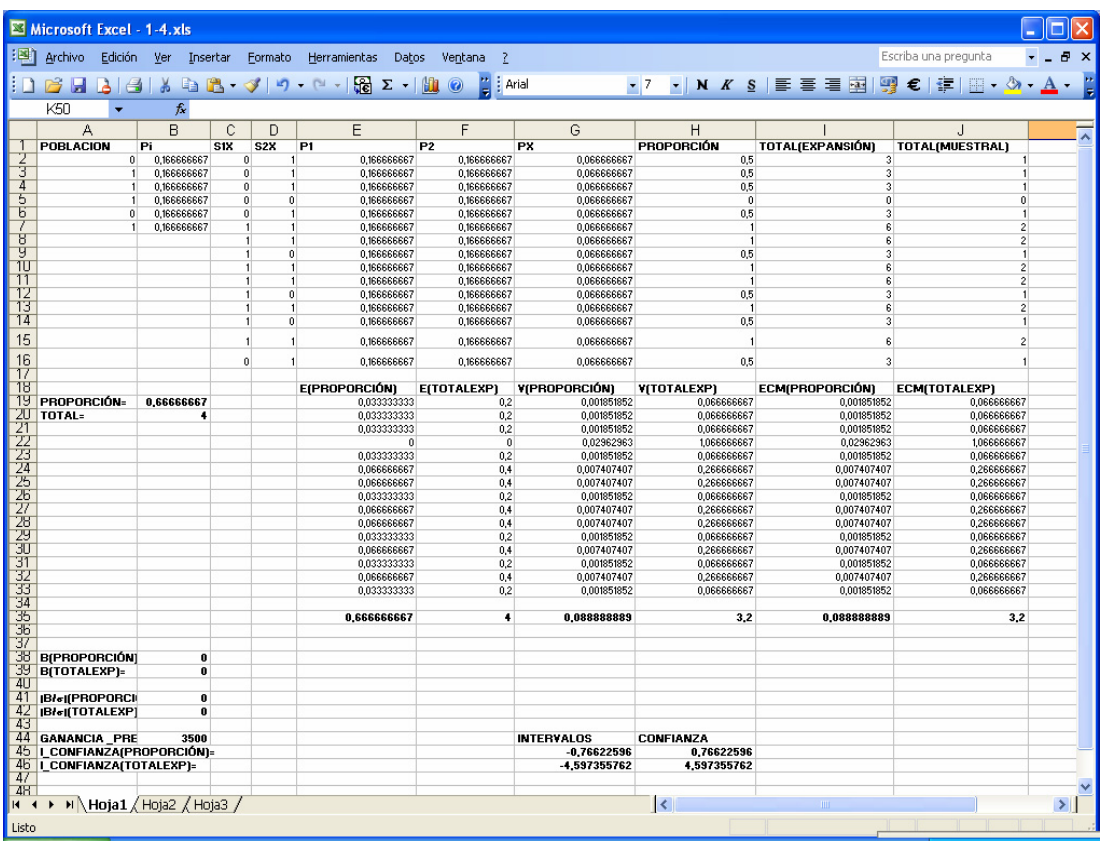
Arial 7

K50

	A	B	C	D	E	F	G	H	I	J
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	PROPORCIÓN	TOTAL(EXPANSIÓN)	TOTAL(MUESTRA
2	0	=1/6	0	1	=1/6	=1/6	=E2*F2/(1-E2)+E2*F2/(1-F2)	=(C2+D2)/2	=6*H2	=2*H2
3	1	=1/6	0	1	=1/6	=1/6	=E3*F3/(1-E3)+E3*F3/(1-F3)	=(C3+D3)/2	=6*H3	=2*H3
4	1	=1/6	0	1	=1/6	=1/6	=E4*F4/(1-E4)+E4*F4/(1-F4)	=(C4+D4)/2	=6*H4	=2*H4
5	1	=1/6	0	0	=1/6	=1/6	=E5*F5/(1-E5)+E5*F5/(1-F5)	=(C5+D5)/2	=6*H5	=2*H5
6	0	=1/6	0	1	=1/6	=1/6	=E6*F6/(1-E6)+E6*F6/(1-F6)	=(C6+D6)/2	=6*H6	=2*H6
7	1	=1/6	1	1	=1/6	=1/6	=E7*F7/(1-E7)+E7*F7/(1-F7)	=(C7+D7)/2	=6*H7	=2*H7
8			1	1	=1/6	=1/6	=E8*F8/(1-E8)+E8*F8/(1-F8)	=(C8+D8)/2	=6*H8	=2*H8
9			1	0	=1/6	=1/6	=E9*F9/(1-E9)+E9*F9/(1-F9)	=(C9+D9)/2	=6*H9	=2*H9
10			1	1	=1/6	=1/6	=E10*F10/(1-E10)+E10*F10/(1-F10)	=(C10+D10)/2	=6*H10	=2*H10
11			1	1	=1/6	=1/6	=E11*F11/(1-E11)+E11*F11/(1-F11)	=(C11+D11)/2	=6*H11	=2*H11
12			1	0	=1/6	=1/6	=E12*F12/(1-E12)+E12*F12/(1-F12)	=(C12+D12)/2	=6*H12	=2*H12
13			1	1	=1/6	=1/6	=E13*F13/(1-E13)+E13*F13/(1-F13)	=(C13+D13)/2	=6*H13	=2*H13
14			1	0	=1/6	=1/6	=E14*F14/(1-E14)+E14*F14/(1-F14)	=(C14+D14)/2	=6*H14	=2*H14
15			1	1	=1/6	=1/6	=E15*F15/(1-E15)+E15*F15/(1-F15)	=(C15+D15)/2	=6*H15	=2*H15
16			0	1	=1/6	=1/6	=E16*F16/(1-E16)+E16*F16/(1-F16)	=(C16+D16)/2	=6*H16	=2*H16
17										
18					E(PROPORCIÓN)	E(TOTALEXP)	V(PROPORCIÓN)	V(TOTALEXP)	ECM(PROPORCIÓN)	ECM(TOTALEXP)
19	PROPORCIÓN=	=PROMEDIO(A2:A7)			=G2*H2	=G2*I2	=G2*(H2-$E$35)^2	=G2*(I2-$F$35)^2	=G2*(H2-$B$19)^2	=G2*(I2-$B$20)^2
20	TOTAL=	=6*B19			=G3*H3	=G3*I3	=G3*(H3-$E$35)^2	=G3*(I3-$F$35)^2	=G3*(H3-$B$19)^2	=G3*(I3-$B$20)^2
21					=G4*H4	=G4*I4	=G4*(H4-$E$35)^2	=G4*(I4-$F$35)^2	=G4*(H4-$B$19)^2	=G4*(I4-$B$20)^2
22					=G5*H5	=G5*I5	=G5*(H5-$E$35)^2	=G5*(I5-$F$35)^2	=G5*(H5-$B$19)^2	=G5*(I5-$B$20)^2
23					=G6*H6	=G6*I6	=G6*(H6-$E$35)^2	=G6*(I6-$F$35)^2	=G6*(H6-$B$19)^2	=G6*(I6-$B$20)^2
24					=G7*H7	=G7*I7	=G7*(H7-$E$35)^2	=G7*(I7-$F$35)^2	=G7*(H7-$B$19)^2	=G7*(I7-$B$20)^2
25					=G8*H8	=G8*I8	=G8*(H8-$E$35)^2	=G8*(I8-$F$35)^2	=G8*(H8-$B$19)^2	=G8*(I8-$B$20)^2
26					=G9*H9	=G9*I9	=G9*(H9-$E$35)^2	=G9*(I9-$F$35)^2	=G9*(H9-$B$19)^2	=G9*(I9-$B$20)^2
27					=G10*H10	=G10*I10	=G10*(H10-$E$35)^2	=G10*(I10-$F$35)^2	=G10*(H10-$B$19)^2	=G10*(I10-$B$20)^2
28					=G11*H11	=G11*I11	=G11*(H11-$E$35)^2	=G11*(I11-$F$35)^2	=G11*(H11-$B$19)^2	=G11*(I11-$B$20)^2
29					=G12*H12	=G12*I12	=G12*(H12-$E$35)^2	=G12*(I12-$F$35)^2	=G12*(H12-$B$19)^2	=G12*(I12-$B$20)^2
30					=G13*H13	=G13*I13	=G13*(H13-$E$35)^2	=G13*(I13-$F$35)^2	=G13*(H13-$B$19)^2	=G13*(I13-$B$20)^2
31					=G14*H14	=G14*I14	=G14*(H14-$E$35)^2	=G14*(I14-$F$35)^2	=G14*(H14-$B$19)^2	=G14*(I14-$B$20)^2
32					=G15*H15	=G15*I15	=G15*(H15-$E$35)^2	=G15*(I15-$F$35)^2	=G15*(H15-$B$19)^2	=G15*(I15-$B$20)^2
33					=G16*H16	=G16*I16	=G16*(H16-$E$35)^2	=G16*(I16-$F$35)^2	=G16*(H16-$B$19)^2	=G16*(I16-$B$20)^2
34										
35					=SUMA(E19:E33)	=SUMA(F19:F33)	=SUMA(G19:G33)	=SUMA(H19:H33)	=SUMA(I19:I33)	=SUMA(J19:J33)
36										
37										
38	B(PROPORCIÓN))=	=E35-B19								
39	B(TOTALEXP)=	=F35-B20								
40										
41	IB/e|(PROPORCIÓN)=	=ABS(B38/RAIZ(G35))								
42	IB/e|(TOTALEXP)=	=ABS(B39/RAIZ(H35))								
43										
44	GANANCIA _PRECIS	=(J35/I35-1)*100					INTERVALOS	CONFIANZA		
45	I_CONFIANZA(PROP						=$H$5-2,57*RAIZ($G$35)	=$H$5+2,57*RAIZ($G$35)		
46	I_CONFIANZA(TOTAI						=$I$5-2,57*RAIZ($H$35)	=$I$5+2,57*RAIZ($H$35)		
47										
48										

Hoja1 Hoja2 Hoja3

Listo



Microsoft Excel - 1-4.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

Arial 7

K50

	A	B	C	D	E	F	G	H	I	J
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	PROPORCIÓN	TOTAL(EXPANSIÓN)	TOTAL(MUESTRAL)
2	0	0,166666667	0	1	0,166666667	0,166666667	0,066666667	0,5	3	1
3	1	0,166666667	0	1	0,166666667	0,166666667	0,066666667	0,5	3	1
4	1	0,166666667	0	1	0,166666667	0,166666667	0,066666667	0,5	3	1
5	1	0,166666667	0	0	0,166666667	0,166666667	0,066666667	0	0	0
6	0	0,166666667	0	1	0,166666667	0,166666667	0,066666667	0,5	3	1
7	1	0,166666667	1	1	0,166666667	0,166666667	0,066666667	1	6	2
8			1	1	0,166666667	0,166666667	0,066666667	1	6	2
9			1	0	0,166666667	0,166666667	0,066666667	0,5	3	1
10			1	1	0,166666667	0,166666667	0,066666667	1	6	2
11			1	1	0,166666667	0,166666667	0,066666667	1	6	2
12			1	0	0,166666667	0,166666667	0,066666667	0,5	3	1
13			1	1	0,166666667	0,166666667	0,066666667	1	6	2
14			1	0	0,166666667	0,166666667	0,066666667	0,5	3	1
15			1	1	0,166666667	0,166666667	0,066666667	1	6	2
16			0	1	0,166666667	0,166666667	0,066666667	0,5	3	1
17										
18					E(PROPORCIÓN)	E(TOTALEXP)	V(PROPORCIÓN)	V(TOTALEXP)	ECM(PROPORCIÓN)	ECM(TOTALEXP)
19	PROPORCIÓN=	0,66666667			0,033333333	0,2	0,001851852	0,066666667	0,001851852	0,066666667
20	TOTAL=	4			0,033333333	0,2	0,001851852	0,066666667	0,001851852	0,066666667
21					0,033333333	0,2	0,001851852	0,066666667	0,001851852	0,066666667
22					0	0	0,02962963	1,066666667	0,02962963	1,066666667
23					0,033333333	0,2	0,001851852	0,066666667	0,001851852	0,066666667
24					0,066666667	0,4	0,007407407	0,266666667	0,007407407	0,266666667
25					0,066666667	0,4	0,007407407	0,266666667	0,007407407	0,266666667
26					0,033333333	0,2	0,001851852	0,066666667	0,001851852	0,066666667
27					0,066666667	0,4	0,007407407	0,266666667	0,007407407	0,266666667
28					0,066666667	0,4	0,007407407	0,266666667	0,007407407	0,266666667
29					0,033333333	0,2	0,001851852	0,066666667	0,001851852	0,066666667
30					0,066666667	0,4	0,007407407	0,266666667	0,007407407	0,266666667
31					0,033333333	0,2	0,001851852	0,066666667	0,001851852	0,066666667
32					0,066666667	0,4	0,007407407	0,266666667	0,007407407	0,266666667
33					0,033333333	0,2	0,001851852	0,066666667	0,001851852	0,066666667
34										
35					0,666666667	4	0,088888889	3,2	0,088888889	3,2
36										
37										
38	B(PROPORCIÓN]	0								
39	B(TOTALEXP]=	0								
40										
41	IB/e|(PROPORCI	0								
42	IB/e|(TOTALEXP]	0								
43										
44	GANANCIA _PRE	3500					INTERVALOS	CONFIANZA		
45	I_CONFIANZA(PROPORCIÓN)=						-0,76622596	0,76622596		
46	I_CONFIANZA(TOTALEXP)=						-4,597355762	4,597355762		
47										
48										

Hoja1 Hoja2 Hoja3

Listo

1.5. Supongamos que los gastos X y los ingresos Y de una empresa a lo largo de los 6 últimos meses fueron los siguientes:

X	3	4	2	2,5	3,5	4,5
Y	6	7	4	5	6,5	8

Se extraen muestras aleatorias simples de dos meses sin reposición y con probabilidades iguales y se pide:

1) Distribución en el muestreo de los estimadores por analogía del gasto total y del estimador por analogía de la proporción que significan los gastos en los ingresos (razón de gastos totales sobre ingresos totales). ¿Qué estimador es mejor? Calcular la ganancia en precisión y expresar los resultados en términos de intervalos de confianza al 95% basados en la muestra de mayor total.

2) Distribución en el muestreo de los estimadores del gasto total siguientes:

   Estimador de expansión del gasto total.
   Proporción de los gastos en los ingresos por el ingreso total poblacional

   ¿Qué estimador es mejor?

Como se trata de muestreo aleatorio sin reposición en el que se supone que el orden de colocación de los elementos en las muestras de tamaño 2 no interviene, el número de muestras posibles, tanto para X como para Y, será $\binom{6}{2}$=15.

Al tratarse de muestreo aleatorio sin reposición y probabilidades iguales, las probabilidades iniciales de selección de los elementos de la población para la muestra valdrán $P(u_i) = 1/6$, i = 1, ..., 6 y la probabilidad de cualquier muestra, tanto para *X* como para *Y*, puede hallarse mediante:

$$P_X = P(u_i,u_j) = P(u_i)P(u_j/u_i) + P(u_j)P(u_i/u_j) = P(u_i)P(u_j)/(1-P(u_i)) + P(u_j)P(u_i)/(1-P(u_j))) =$$

$$(1/6^2)/(1-1/6) + (1/6^2)/(1-1/6) = 2(1/6^2)/(1-1/6) = 1/15 = 0{,}066666$$

Se observa que las probabilidades de las muestras serán todas iguales a 1/15. Luego estamos ante un método de selección con probabilidades iguales y muestras equiprobables.

A continuación se presenta la tabla que contiene, para *X* y para *Y*, el espacio muestral, las probabilidades de las muestras y la distribución de los estimadores.

Denominamos *GTOTAL* al estimador por analogía del gasto total (total muestral del gasto) y *RAZÓN* al estimador por analogía de la proporción que significan los gastos en los ingresos (total muestral del gasto entre total muestral del ingreso). Se tendrá presente que el estimador expandido del gasto total es el producto del tamaño poblacional por la media muestral del gasto ($GTOTALEXP = 6(GTOTAL/2) = 3GTOTAL$) y que la proporción de los gastos en los ingresos por el ingreso total poblacional es $TOTAL = (36{,}5)RAZÓN$. En los estimadores, para las cuatro primeras filas de la tabla se indican todas las operaciones y para el resto de las filas las operaciones son similares y se indican sólo los resultados.

S1_X	S2_X	S1_Y	S2_Y	P=PX=PY	GTOTAL	RAZÓN	GTOTALEXP (3*GTOTAL)	TOTAL (36,5*RAZÓN)
3	4	6	7	1/15	7=3+4	0,53=(3+4)/(6+7)	21=3*7	19,65=36,5*0,53
3	2	6	4	1/15	5=3+2	0,5=(3+2)/(6+4)	15=3*5	18,25=36,5*0,5
3	2,5	6	5	1/15	5,5=3+2,5	0,5=(3+2,5)/(6+5)	16,5=3*5,5	18,25=36,5*0,5
3	3,5	6	6,5	1/15	6,5=3+3,5	0,52=(3+3,5)/(6+6,5)	19,5=3*6,5	18,98=36,5*0,52
3	4,5	6	8	1/15	7,5	0,535714286	22,5	19,55357143
4	2	7	4	1/15	6	0,545454545	18	19,90909091
4	2,5	7	5	1/15	6,5	0,541666667	19,5	19,77083333
4	3,5	7	6,5	1/15	7,5	0,555555556	22,5	20,27777778
4	4,5	7	8	1/15	8,5	0,566666667	25,5	20,68333333
2	2,5	4	5	1/15	4,5	0,5	13,5	18,25
2	3,5	4	6,5	1/15	5,5	0,523809524	16,5	19,11904762
2	4,5	4	8	1/15	6,5	0,541666667	19,5	19,77083333
2,5	3,5	5	6,5	1/15	6	0,52173913	18	19,04347826
2,5	4,5	5	8	1/15	7	0,538461538	21	19,65384615
3,5	4,5	6,5	8	1/15	8	0,551724138	24	20,13793103

Una vez conocida la distribución de probabilidad en el muestreo de los estimadores compararemos en primer lugar el estimador *GTOTAL* (que estima el gasto total poblacional θ_1) y *RAZÓN* (que estima la proporción de los gastos totales sobre los ingresos totales en la población θ_2).

Para comprobar la insesgadez, hallamos la esperanza matemática de los estimadores tal y como se indica a continuación:

$$E(GTOTAL) = \sum_{i=1}^{15} GTOTAL_i P_i = 7 \cdot \frac{1}{15} + 5 \cdot \frac{1}{15} + \cdots + 8 \cdot \frac{1}{15} = 6{,}5 \neq 19{,}5 = \theta_1$$

$$E(RAZÓN) = \sum_{i=1}^{15} RAZÓN_i P_i = 0{,}53 \cdot \frac{1}{15} + 0{,}5 \cdot \frac{1}{15} + \cdots + 0{,}55 \cdot \frac{1}{15} = 0{,}53206 \neq 0{,}53424 = \theta_2$$

Para calcular los sesgos se observa que $B(GTOTAL) = E(GTOTAL) - \theta_1 = 6{,}5 - 19{,}5 = -13$ y $B(RAZÓN) = E(RAZÓN) - \theta_2 = 0{,}53206 - 0{,}53424 = -0{,}00218$. A continuación se calculan las varianzas de los estimadores.

$$V(GTOTAL) = \sum_{i=1}^{15} \left(GTOTAL_i - E(GTOTAL)\right)^2 P_i = (7 - 6{,}5)^2 \cdot \frac{1}{15} + \cdots + (8 - 6{,}5)^2 \cdot \frac{1}{15} = 1{,}1666$$

$$V(RAZÓN) = \sum_{i=1}^{15} \left(RAZÓN_i - E(RAZÓN)\right)^2 P_i = (0{,}53 - 0{,}532)^2 \cdot \frac{1}{15} + \cdots + (0{,}55 - 0{,}532)^2 \cdot \frac{1}{15} = 0{,}000399$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(GTOTAL) = \sqrt{1{,}1666} = 1{,}08, \quad \sigma(RAZÓN) = \sqrt{0{,}000399} = 0{,}0199$$

Como $|B(GTOTAL)/\sigma(GTOTAL)| = 12{,}03 > 1/10$, el sesgo del estimador *GTOTAL* no es despreciable y como $|B(RAZÓN)/\sigma(RAZÓN)| = 0{,}1 \leq 1/10$, el sesgo de *RAZÓN* es despreciable y a todos los efectos este estimador es insesgado. Al comparar *RAZÓN* con *GTOTAL* tenemos un estimador sesgado y el otro insesgado. La comparación debe hacerse a través de los errores cuadráticos medios. Tenemos:

$$ECM(GTOTAL) = \sum_{i=1}^{15} \left(GTOTAL_i - \theta_1\right)^2 P_i = (7-19{,}5)^2 \cdot \frac{1}{15} + \cdots + (8-19{,}5)^2 \cdot \frac{1}{15} = 170{,}166$$

$$ECM(RAZÓN) = V(RAZÓN) = 0{,}00399$$

Como el estimador *RAZÓN* es insesgado, su varianza coincide con su error cuadrático medio, luego su precisión se mide a través de la varianza. De esta forma, el estimador *RAZÓN* para estimar θ_2 es más preciso que el estimador *GTOTAL* para estimar θ_1 por tener menor error cuadrático medio.

La ganancia en precisión de *RAZÓN* respecto de *GTOTAL s*e cuantifica como sigue:

$$GP = (EMC\ (GTOTAL)/EMC\ (RAZÓN) - 1)100 = (170{,}166/0{,}00399\text{-}1)100 = 42045172{,}1\%$$

El intervalo de confianza para *GTOTAL* (sesgado) basado en la muestra de mayor total al 95% es el siguiente:

$$[\hat{\theta} - \lambda_\alpha \sigma(\hat{\theta}) - |B(\hat{\theta})|, \hat{\theta} + \lambda_\alpha \sigma(\hat{\theta}) + |B(\hat{\theta})|] = [8{,}5 - 1{,}96 \cdot 1{,}08 - 13,\quad 8{,}5 + 1{,}96 \cdot 1{,}08 + 13] = [-6.61, 23.61]$$

Suponiendo normalidad el intervalo de confianza para *RAZÓN* (insesgado) al 95% basado en la muestra de mayor total sería:

$$[\hat{\theta} - \lambda_\alpha \sigma(\hat{\theta}), \hat{\theta} + \lambda_\alpha \sigma(\hat{\theta})] = [0{,}566 - 1{,}96 \cdot 0{,}0199,\quad 0{,}566 + 1{,}96 \cdot 0{,}0199] = [0.527, 0.605]$$

Se observa que el intervalo de confianza relativo a *RAZÓN* es bastante más estrecho (más preciso) que el relativo a *GTOTAL*. Esta fuerte diferencia de anchuras de intervalos está en línea con la cuantía tan fuerte de ganancia en precisión de *RAZÓN* sobre *GTOTAL*.

Para comparar los estimadores del gasto total *GTOTAEXP* y *TOTAL*, observamos que *GTOTALEXP* = 3*GTOTAL* y *TOTAL* = (36,5)*RAZÓN*. Tenemos:

$E(GTOTALEXP) = 3E(GTOTAL) = 3(6{,}5) = 19{,}5 = \theta_1$

$E(TOTAL) = (36{,}5)E(RAZÓN) = (36{,}5)(0{,}532) = 19{,}42 \approx \theta_1$

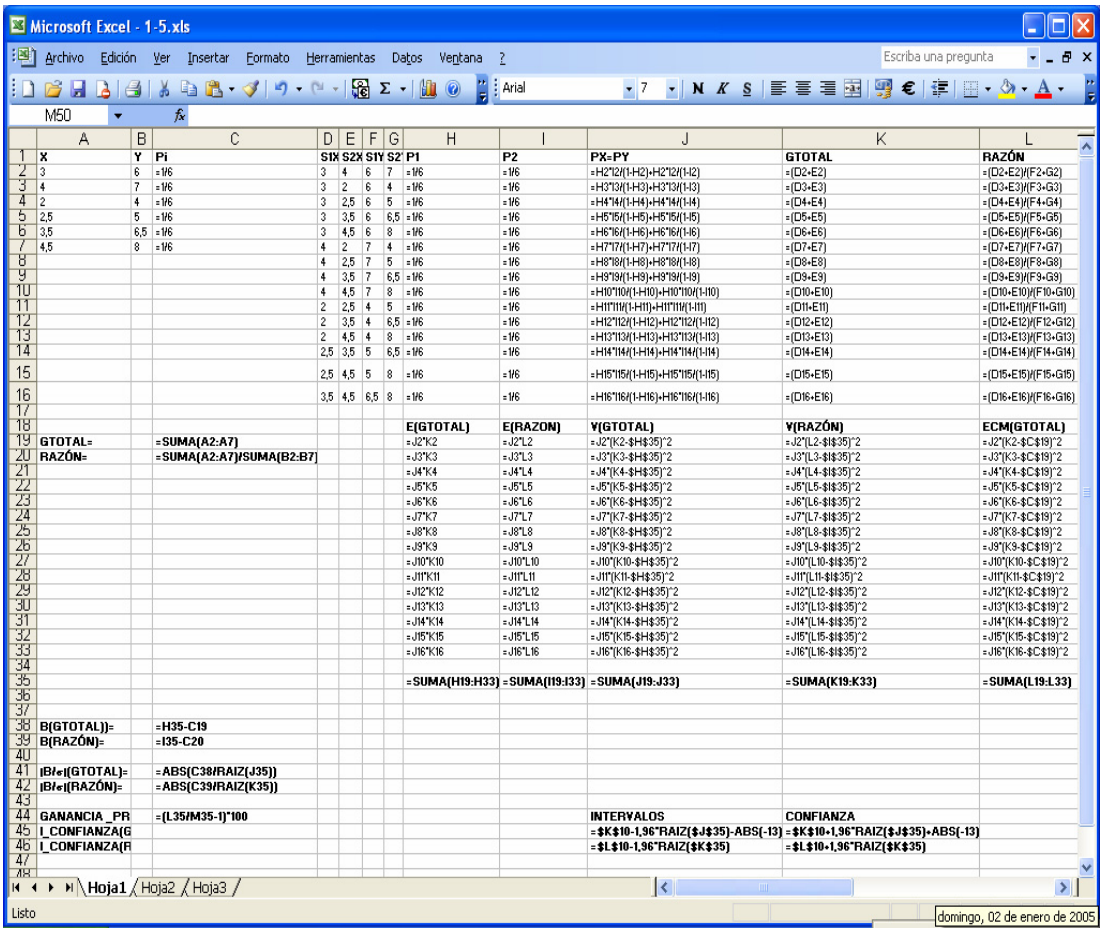
$V(GTOTALEXP) = 9V(GTOTAL) = 9(1{,}166) = 10{,}5$

$V(TOTAL) = (36{,}5^2)V(RAZÓN) = (36{,}5^2)0{,}000399 = 0{,}539$

Los dos estimadores han resultado ser insesgados, con lo que será más preciso el que tenga menor varianza; es decir, *TOTAL* es más preciso que *GTOTALEXP*.

A continuación se presentan los cálculos anteriores automatizados a través de Excel. Las hoja de Excel con las fórmulas se ha dividido en dos trozos debido a la extensión de los cálculos necesarios.

A continuación de las dos hojas de fórmulas se presenta la hoja de resultados.



Microsoft Excel - 1-5.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

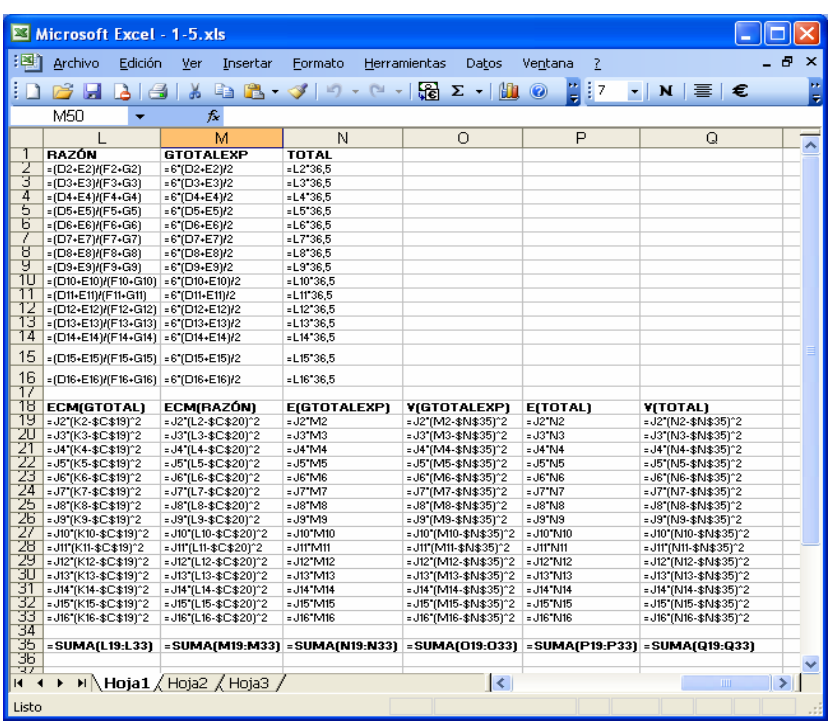
M50

	A	B	C	D	E	F	G	H	I	J	K	L
1	X	Y	Pi	S1X	S2X	S1Y	S2Y	P1	P2	PX-PY	GTOTAL	RAZÓN
2	3	6	=1/6	3	4	6	7	=1/6	=1/6	=H2*I2/(1-H2)+H2*I2/(1-I2)	=D2+E2	=D2+E2)/(F2+G2)
3	4	7	=1/6	3	2	6	4	=1/6	=1/6	=H3*I3/(1-H3)+H3*I3/(1-I3)	=D3+E3	=D3+E3)/(F3+G3)
4	2	4	=1/6	3	2,5	6	5	=1/6	=1/6	=H4*I4/(1-H4)+H4*I4/(1-I4)	=D4+E4	=D4+E4)/(F4+G4)
5	2,5	5	=1/6	3	3,5	6	6,5	=1/6	=1/6	=H5*I5/(1-H5)+H5*I5/(1-I5)	=D5+E5	=D5+E5)/(F5+G5)
6	3,5	6,5	=1/6	3	4,5	6	8	=1/6	=1/6	=H6*I6/(1-H6)+H6*I6/(1-I6)	=D6+E6	=D6+E6)/(F6+G6)
7	4,5	8	=1/6	4	2	7	4	=1/6	=1/6	=H7*I7/(1-H7)+H7*I7/(1-I7)	=D7+E7	=D7+E7)/(F7+G7)
8				4	2,5	7	5	=1/6	=1/6	=H8*I8/(1-H8)+H8*I8/(1-I8)	=D8+E8	=D8+E8)/(F8+G8)
9				4	3,5	7	6,5	=1/6	=1/6	=H9*I9/(1-H9)+H9*I9/(1-I9)	=D9+E9	=D9+E9)/(F9+G9)
10				4	4,5	7	8	=1/6	=1/6	=H10*I10/(1-H10)+H10*I10/(1-I10)	=D10+E10	=D10+E10)/(F10+G10)
11				2	2,5	4	5	=1/6	=1/6	=H11*I11/(1-H11)+H11*I11/(1-I11)	=D11+E11	=D11+E11)/(F11+G11)
12				2	3,5	4	6,5	=1/6	=1/6	=H12*I12/(1-H12)+H12*I12/(1-I12)	=D12+E12	=D12+E12)/(F12+G12)
13				2	4,5	4	8	=1/6	=1/6	=H13*I13/(1-H13)+H13*I13/(1-I13)	=D13+E13	=D13+E13)/(F13+G13)
14				2,5	3,5	5	6,5	=1/6	=1/6	=H14*I14/(1-H14)+H14*I14/(1-I14)	=D14+E14	=D14+E14)/(F14+G14)
15				2,5	4,5	5	8	=1/6	=1/6	=H15*I15/(1-H15)+H15*I15/(1-I15)	=D15+E15	=D15+E15)/(F15+G15)
16				3,5	4,5	6,5	8	=1/6	=1/6	=H16*I16/(1-H16)+H16*I16/(1-I16)	=D16+E16	=D16+E16)/(F16+G16)
17												
18								E(GTOTAL)	E(RAZON)	V(GTOTAL)	V(RAZÓN)	ECM(GTOTAL)
19	GTOTAL=		=SUMA(A2:A7)					=J2*K2	=J2*L2	=J2*(K2-$H$35)^2	=J2*(L2-$I$35)^2	=J2*(K2-$C$19)^2
20	RAZÓN=		=SUMA(A2:A7)/SUMA(B2:B7)					=J3*K3	=J3*L3	=J3*(K3-$H$35)^2	=J3*(L3-$I$35)^2	=J3*(K3-$C$19)^2
21								=J4*K4	=J4*L4	=J4*(K4-$H$35)^2	=J4*(L4-$I$35)^2	=J4*(K4-$C$19)^2
22								=J5*K5	=J5*L5	=J5*(K5-$H$35)^2	=J5*(L5-$I$35)^2	=J5*(K5-$C$19)^2
23								=J6*K6	=J6*L6	=J6*(K6-$H$35)^2	=J6*(L6-$I$35)^2	=J6*(K6-$C$19)^2
24								=J7*K7	=J7*L7	=J7*(K7-$H$35)^2	=J7*(L7-$I$35)^2	=J7*(K7-$C$19)^2
25								=J8*K8	=J8*L8	=J8*(K8-$H$35)^2	=J8*(L8-$I$35)^2	=J8*(K8-$C$19)^2
26								=J9*K9	=J9*L9	=J9*(K9-$H$35)^2	=J9*(L9-$I$35)^2	=J9*(K9-$C$19)^2
27								=J10*K10	=J10*L10	=J10*(K10-$H$35)^2	=J10*(L10-$I$35)^2	=J10*(K10-$C$19)^2
28								=J11*K11	=J11*L11	=J11*(K11-$H$35)^2	=J11*(L11-$I$35)^2	=J11*(K11-$C$19)^2
29								=J12*K12	=J12*L12	=J12*(K12-$H$35)^2	=J12*(L12-$I$35)^2	=J12*(K12-$C$19)^2
30								=J13*K13	=J13*L13	=J13*(K13-$H$35)^2	=J13*(L13-$I$35)^2	=J13*(K13-$C$19)^2
31								=J14*K14	=J14*L14	=J14*(K14-$H$35)^2	=J14*(L14-$I$35)^2	=J14*(K14-$C$19)^2
32								=J15*K15	=J15*L15	=J15*(K15-$H$35)^2	=J15*(L15-$I$35)^2	=J15*(K15-$C$19)^2
33								=J16*K16	=J16*L16	=J16*(K16-$H$35)^2	=J16*(L16-$I$35)^2	=J16*(K16-$C$19)^2
34												
35								=SUMA(H19:H33)	=SUMA(I19:I33)	=SUMA(J19:J33)	=SUMA(K19:K33)	=SUMA(L19:L33)
36												
37												
38	B(GTOTAL))=		=H35-C19									
39	B(RAZÓN)=		=I35-C20									
40												
41	IBIasI(GTOTAL)=		=ABS(C38/RAIZ(J35))									
42	IBIasI(RAZÓN)=		=ABS(C39/RAIZ(K35))									
43												
44	GANANCIA_PR		=(L35/M35-1)*100							INTERVALOS	CONFIANZA	
45	I_CONFIANZA(G									=$K$10-1,96*RAIZ($J$35)-ABS(-13)	=$K$10+1,96*RAIZ($J$35)+ABS(-13)	
46	I_CONFIANZA(R									=$L$10-1,96*RAIZ($K$35)	=$L$10+1,96*RAIZ($K$35)	
47												

Hoja1 Hoja2 Hoja3

Listo

domingo, 02 de enero de 2005



Microsoft Excel - 1-5.xls

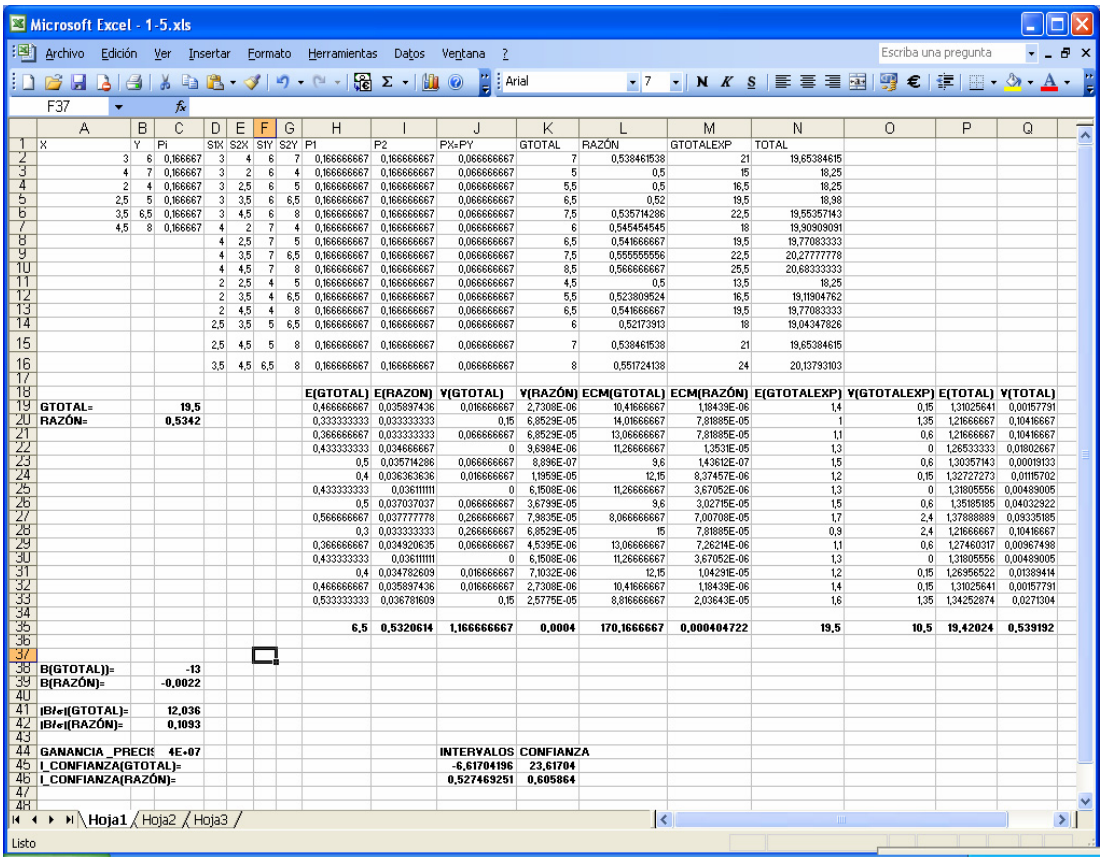
Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

M50

	L	M	N	O	P	Q
1	RAZÓN	GTOTALEXP	TOTAL			
2	=(D2+E2)/(F2+G2)	=6*(D2+E2)/2	=L2*36,5			
3	=(D3+E3)/(F3+G3)	=6*(D3+E3)/2	=L3*36,5			
4	=(D4+E4)/(F4+G4)	=6*(D4+E4)/2	=L4*36,5			
5	=(D5+E5)/(F5+G5)	=6*(D5+E5)/2	=L5*36,5			
6	=(D6+E6)/(F6+G6)	=6*(D6+E6)/2	=L6*36,5			
7	=(D7+E7)/(F7+G7)	=6*(D7+E7)/2	=L7*36,5			
8	=(D8+E8)/(F8+G8)	=6*(D8+E8)/2	=L8*36,5			
9	=(D9+E9)/(F9+G9)	=6*(D9+E9)/2	=L9*36,5			
10	=(D10+E10)/(F10+G10)	=6*(D10+E10)/2	=L10*36,5			
11	=(D11+E11)/(F11+G11)	=6*(D11+E11)/2	=L11*36,5			
12	=(D12+E12)/(F12+G12)	=6*(D12+E12)/2	=L12*36,5			
13	=(D13+E13)/(F13+G13)	=6*(D13+E13)/2	=L13*36,5			
14	=(D14+E14)/(F14+G14)	=6*(D14+E14)/2	=L14*36,5			
15	=(D15+E15)/(F15+G15)	=6*(D15+E15)/2	=L15*36,5			
16	=(D16+E16)/(F16+G16)	=6*(D16+E16)/2	=L16*36,5			
17						
18	ECM(GTOTAL)	ECM(RAZÓN)	E(GTOTALEXP)	V(GTOTALEXP)	E(TOTAL)	V(TOTAL)
19	=J2*(K2-$C$19)^2	=J2*(L2-$C$20)^2	=J2*M2	=J2*(M2-$N$35)^2	=J2*N2	=J2*(N2-$N$35)^2
20	=J3*(K3-$C$19)^2	=J3*(L3-$C$20)^2	=J3*M3	=J3*(M3-$N$35)^2	=J3*N3	=J3*(N3-$N$35)^2
21	=J4*(K4-$C$19)^2	=J4*(L4-$C$20)^2	=J4*M4	=J4*(M4-$N$35)^2	=J4*N4	=J4*(N4-$N$35)^2
22	=J5*(K5-$C$19)^2	=J5*(L5-$C$20)^2	=J5*M5	=J5*(M5-$N$35)^2	=J5*N5	=J5*(N5-$N$35)^2
23	=J6*(K6-$C$19)^2	=J6*(L6-$C$20)^2	=J6*M6	=J6*(M6-$N$35)^2	=J6*N6	=J6*(N6-$N$35)^2
24	=J7*(K7-$C$19)^2	=J7*(L7-$C$20)^2	=J7*M7	=J7*(M7-$N$35)^2	=J7*N7	=J7*(N7-$N$35)^2
25	=J8*(K8-$C$19)^2	=J8*(L8-$C$20)^2	=J8*M8	=J8*(M8-$N$35)^2	=J8*N8	=J8*(N8-$N$35)^2
26	=J9*(K9-$C$19)^2	=J9*(L9-$C$20)^2	=J9*M9	=J9*(M9-$N$35)^2	=J9*N9	=J9*(N9-$N$35)^2
27	=J10*(K10-$C$19)^2	=J10*(L10-$C$20)^2	=J10*M10	=J10*(M10-$N$35)^2	=J10*N10	=J10*(N10-$N$35)^2
28	=J11*(K11-$C$19)^2	=J11*(L11-$C$20)^2	=J11*M11	=J11*(M11-$N$35)^2	=J11*N11	=J11*(N11-$N$35)^2
29	=J12*(K12-$C$19)^2	=J12*(L12-$C$20)^2	=J12*M12	=J12*(M12-$N$35)^2	=J12*N12	=J12*(N12-$N$35)^2
30	=J13*(K13-$C$19)^2	=J13*(L13-$C$20)^2	=J13*M13	=J13*(M13-$N$35)^2	=J13*N13	=J13*(N13-$N$35)^2
31	=J14*(K14-$C$19)^2	=J14*(L14-$C$20)^2	=J14*M14	=J14*(M14-$N$35)^2	=J14*N14	=J14*(N14-$N$35)^2
32	=J15*(K15-$C$19)^2	=J15*(L15-$C$20)^2	=J15*M15	=J15*(M15-$N$35)^2	=J15*N15	=J15*(N15-$N$35)^2
33	=J16*(K16-$C$19)^2	=J16*(L16-$C$20)^2	=J16*M16	=J16*(M16-$N$35)^2	=J16*N16	=J16*(N16-$N$35)^2
34						
35	=SUMA(L19:L33)	=SUMA(M19:M33)	=SUMA(N19:N33)	=SUMA(O19:O33)	=SUMA(P19:P33)	=SUMA(Q19:Q33)
36						

Hoja1 Hoja2 Hoja3

Listo



Microsoft Excel - 1-5.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

F37

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	X	Y	Pi	STX	S2X	STY	S2Y	P1	P2	PX+PY	GTOTAL	RAZÓN	GTOTALEXP	TOTAL			
2	3	6	0,166667	3	4	6	7	0,166666667	0,166666667	0,066666667	7	0,538461538	21	19,65384615			
3	4	7	0,166667	3	2	6	4	0,166666667	0,166666667	0,066666667	5	0,5	15	18,25			
4	2	4	0,166667	3	2,5	6	5	0,166666667	0,166666667	0,066666667	5,5	0,5	16,5	18,25			
5	2,5	5	0,166667	3	3,5	6	6,5	0,166666667	0,166666667	0,066666667	6,5	0,52	19,5	18,98			
6	3,5	6,5	0,166667	3	4,5	6	8	0,166666667	0,166666667	0,066666667	7,5	0,535714286	22,5	19,55357143			
7	4,5	8	0,166667	4	2	7	4	0,166666667	0,166666667	0,066666667	6	0,545454545	18	19,90909091			
8				4	2,5	7	5	0,166666667	0,166666667	0,066666667	6,5	0,541666667	19,5	19,77083333			
9				4	3,5	7	6,5	0,166666667	0,166666667	0,066666667	7,5	0,555555556	22,5	20,27777778			
10				4	4,5	7	8	0,166666667	0,166666667	0,066666667	8,5	0,566666667	25,5	20,68333333			
11				2	2,5	4	5	0,166666667	0,166666667	0,066666667	4,5	0,5	13,5	18,25			
12				2	3,5	4	6,5	0,166666667	0,166666667	0,066666667	5,5	0,523809524	16,5	19,11904762			
13				2	4,5	4	8	0,166666667	0,166666667	0,066666667	6,5	0,541666667	19,5	19,77083333			
14				2,5	3,5	5	6,5	0,166666667	0,166666667	0,066666667	6	0,52173913	18	19,04347826			
15				2,5	4,5	5	8	0,166666667	0,166666667	0,066666667	7	0,538461538	21	19,65384615			
16				3,5	4,5	6,5	8	0,166666667	0,166666667	0,066666667	8	0,551724138	24	20,13793103			
17																	
18								E(GTOTAL)	E(RAZON)	V(GTOTAL)	V(RAZÓN)	ECM(GTOTAL)	ECM(RAZÓN)	E(GTOTALEXP)	V(GTOTALEXP)	E(TOTAL)	V(TOTAL)
19	GTOTAL=		19,5					0,466666667	0,035897436	0,016666667	2,7308E-06	10,41666667	1,18439E-06	1,4	0,15	1,31025641	0,00157791
20	RAZÓN=		0,5342					0,333333333	0,033333333	0,15	6,8529E-05	14,01666667	7,81885E-05	1	1,35	1,21666667	0,10416667
21								0,366666667	0,033333333	0,066666667	6,8529E-05	13,06666667	7,81885E-05	1,1	0,6	1,21666667	0,10416667
22								0,433333333	0,034666667	0	9,6984E-06	11,26666667	1,3531E-05	1,3	0	1,26533333	0,01802667
23								0,5	0,035714286	0,066666667	8,896E-07	9,6	1,43612E-07	1,5	0,6	1,30357143	0,00019133
24								0,4	0,036363636	0,016666667	1,1959E-05	12,15	8,37457E-06	1,2	0,15	1,32727273	0,01115702
25								0,433333333	0,03611111	0	6,1508E-06	11,26666667	3,67052E-06	1,3	0	1,31805556	0,00489005
26								0,5	0,037037037	0,066666667	3,6799E-05	9,6	3,02715E-05	1,5	0,6	1,35185185	0,04032922
27								0,566666667	0,037777778	0,266666667	7,9835E-05	8,066666667	7,00708E-05	1,7	2,4	1,37888889	0,09335185
28								0,3	0,033333333	0,266666667	6,8529E-05	15	7,81885E-05	0,9	2,4	1,21666667	0,10416667
29								0,366666667	0,034920635	0,066666667	4,5395E-06	13,06666667	7,26214E-06	1,1	0,6	1,27460317	0,00967498
30								0,433333333	0,03611111	0	6,1508E-06	11,26666667	3,67052E-06	1,3	0	1,31805556	0,00489005
31								0,4	0,034782609	0,016666667	7,1032E-06	12,15	1,04291E-05	1,2	0,15	1,26956522	0,01389414
32								0,466666667	0,035897436	0,016666667	2,7308E-06	10,41666667	1,18439E-06	1,4	0,15	1,31025641	0,00157791
33								0,533333333	0,036781609	0,15	2,5775E-05	8,816666667	2,03643E-05	1,6	1,35	1,34252874	0,0271304
34																	
35								6,5	0,5320614	1,166666667	0,0004	170,1666667	0,000404722	19,5	10,5	19,42024	0,539192
36																	
37																	
38	B(GTOTAL))=		-13														
39	B(RAZÓN)=		-0,0022														
40																	
41	IBIeI(GTOTAL)=		12,036														
42	IBIeI(RAZÓN)=		0,1093														
43																	
44	GANANCIA_PRECIS		4E+07							INTERVALOS	CONFIANZA						
45	I_CONFIANZA(GTOTAL)=									-6,6170496	23,61704						
46	I_CONFIANZA(RAZÓN)=									0,527469251	0,605864						
47																	
48																	

Hoja1 Hoja2 Hoja3

Listo

1.6. Consideramos una población virtual para simulación formada por 10 individuos agrupados en 4 hogares y cuyos ingresos anuales en miles de euros (variable X) se presentan en la tabla adjunta:

HOGARES →	H1	H2	H3	H4
INGRESOS (X_i) →	1, 2, 3	4, 6	9, 11	2, 2, 5

Se considera un procedimiento de muestreo que consiste en elegir cada hogar con probabilidades proporcionales a sus tamaños. Se considera el estimador T_1 = Ingreso medio de los hogares, para estimar el ingreso medio poblacional, y se considera el estimador T_2 = Ingreso total de los hogares, para estimar el ingreso total poblacional. Se pide:

1) Especificar el espacio muestral relativo a este procedimiento de muestreo y las probabilidades asociadas a las muestras. Hallar también las distribuciones de probabilidad en el muestreo de los estimadores T_1 y T_2. ¿Cuál de ellos es mejor? Razonar la respuesta y cuantificar la ganancia en precisión.

2) Hallar un intervalo de confianza para el ingreso medio al nivel $\alpha = 0{,}002$ basado en el subconjunto de mayor total. Se sabe que $F^{-1}(0{,}999) = 3$, siendo F la función de distribución de una Normal (0,1). Hallar también un intervalo de confianza del 95% para el ingreso total basado en el subconjunto de mayor media. Se sabe que $F^{-1}(0{,}975) = 2$, siendo F la función de distribución de una Normal (0,1).

Como el procedimiento de muestreo es con probabilidades proporcionales a los tamaños M_i de los hogares tenemos que $P_i = kM_i$ $i = 1, 2, 3, 4$ para una constante de proporcionalidad k que se calcula de la forma siguiente:

$$P_i = kM_i \Rightarrow \sum_{i=1}^{4} P_i = k\sum_{i=1}^{4} M_i \Rightarrow 1 = k \cdot 10 \Rightarrow k = 1/10 \Rightarrow \begin{cases} P_1 = 3/10 \\ P_2 = 2/10 = 1/5 \\ P_3 = 2/10 = 1/5 \\ P_4 = 3/10 \end{cases}$$

En el siguiente cuadro se especifican las muestras, sus probabilidades y los valores de los estimadores para cada muestra.

$S(X)$	$P(X)$	T_1	T_2
{1,2,3}	3/10	2	6
{4,6}	1/5	5	10
{9,11}	1/5	10	20
{2,2,5}	3/10	3	9

Las distribuciones de probabilidad de los dos estimadores se calcularán mediante la expresión ya conocida $P^T(\hat{\theta}(X_1, ..., X_n) = t) = \sum_{\{S_i / \hat{\theta}(S_i(X))=t\}} P(S_i)$, de la siguiente forma:

$$T_1 \begin{cases} P^T(T_1 = 2) = P\{1,2,3\} = \dfrac{3}{10} \\ P^T(T_1 = 5) = P\{4,6\} = \dfrac{1}{5} \\ P^T(T_1 = 10) = P\{9,11\} = \dfrac{1}{5} \\ P^T(T_1 = 3) = P\{2,2,5\} = \dfrac{3}{10} \end{cases} \qquad T_2 \begin{cases} P^T(T_2 = 6) = P\{1,2,3\} = \dfrac{3}{10} \\ P^T(T_2 = 10) = P\{4,6\} = \dfrac{1}{5} \\ P^T(T_2 = 20) = P\{9,11\} = \dfrac{1}{5} \\ P^T(T_2 = 9) = P\{2,2,5\} = \dfrac{3}{10} \end{cases}$$

Una vez conocida la distribución de probabilidad en el muestreo de los dos estimadores analizaremos si son insesgados o no. Para ello calculamos en primer lugar los valores de la media poblacional y el total poblacional, que son los parámetros que estamos estimando. Se tiene:

$$\overline{X} = (1 + 2 + 3 + 4 + 6 + 9 + 11 + 2 + 2 + 5)/10 = 45/10$$
$$X = (1 + 2 + 3 + 4 + 6 + 9 + 11 + 2 + 2 + 5) = 45$$

Ahora, para comprobar la insesgadez, hallamos la esperanza matemática de los estimadores tal y como se indica a continuación:

$$E(T_1) = 2 \cdot \frac{3}{10} + 5 \cdot \frac{1}{5} + 10 \cdot \frac{1}{5} + 3 \cdot \frac{3}{10} = 4{,}5 = \overline{X}$$

$$E(T_2) = 6 \cdot \frac{3}{10} + 10 \cdot \frac{1}{5} + 20 \cdot \frac{1}{5} + 9 \cdot \frac{3}{10} = 10{,}5 \neq X = 45$$

El estimador T_1 es insesgado, pero el estimador T_2 es sesgado con sesgo $B(T_2)=E(T_2) - X = 10{,}5 - 45 = -34{,}5$. Las varianzas de los estimadores son:

$$V(T_1) = (2-4{,}5)^2 \cdot \frac{3}{10} + (5-4{,}5)^2 \cdot \frac{1}{5} + (10-4{,}5)^2 \cdot \frac{1}{5} + (3-4{,}5)^2 \cdot \frac{3}{10} = 8{,}65$$

$$V(T_2) = (6-10{,}5)^2 \cdot \frac{3}{10} + (10-10{,}5)^2 \cdot \frac{1}{5} + (20-10{,}5)^2 \cdot \frac{1}{5} + (9-10{,}5)^2 \cdot \frac{3}{10} = 24{,}85$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(T_1) = \sqrt{8{,}65} = 2.94 \quad y \quad \sigma(T_2) = \sqrt{24{,}85} = 4.98$$

Como $|B(T_2)/\sigma(T_2)| = 6.92 > 1/10$, el sesgo del estimador T_2 no es despreciable, y como T_1 es insesgado, la comparación de estimadores ha de hacerse a través del error cuadrático medio. Tenemos:

$$ECM(T_1) = (2-4{,}5)^2 \cdot \frac{3}{10} + (5-4{,}5)^2 \cdot \frac{1}{5} + (10-4{,}5)^2 \cdot \frac{1}{5} + (3-4{,}5)^2 \cdot \frac{3}{10} = 8{,}65$$

$$ECM(T_2) = (6-45)^2 \cdot \frac{3}{10} + (10-45)^2 \cdot \frac{1}{5} + (20-45)^2 \cdot \frac{1}{5} + (9-45)^2 \cdot \frac{3}{10} = 1215{,}1$$

Evidentemente, el mejor estimador es T_1, pues su error cuadrático medio es mucho menor que el de T_2. La ganancia en precisión por usar T_1 en vez de T_2 es:

$$GP=(EMC(T_2)/EMC(T_1) - 1)100 = (1215{,}1/8{,}65\text{-}1)100=13946{,}24\%$$

Para hallar un intervalo de confianza para T_1 (que es insesgado) basado en la muestra de mayor total {9,11}, suponemos primeramente que la población se distribuye normalmente, en cuyo caso se utiliza como intervalo de confianza el siguiente:

$$[\hat{\theta} - \lambda_\alpha \sigma(\hat{\theta}), \hat{\theta} + \lambda_\alpha \sigma(\hat{\theta})] = [10 - 3 \cdot 2.94, 10 + 3 \cdot 2.94] = [1.17, 18.82]$$

Si la población no se distribuye normalmente el intervalo para T_1 es:

$$\left[\hat{\theta} - \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}}, \hat{\theta} + \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}}\right] = \left[10 - \frac{2.94}{\sqrt{0.002}}, 10 + \frac{2.94}{\sqrt{0.002}}\right] = [-55.74, 75.7]$$

Se observa que la longitud del intervalo de confianza cuando no hay normalidad es mucho mayor que en el caso de normalidad, con lo que la estimación es más tosca (peor) en el caso de no normalidad.

Para hallar un intervalo de confianza para T_2 (que es sesgado) basado en la muestra de mayor media {9,11}, realizamos los siguientes cálculos:

$$[\hat{\theta} - \lambda_\alpha \sigma(\hat{\theta}) - |B(\hat{\theta})|, \hat{\theta} + \lambda_\alpha \theta(\hat{\theta}) + |B(\hat{\theta})|] = [20 - 2 \cdot 4.98 - 34.5, 20 + 2 \cdot 4.98 + 34.5] = [-24.47, 64.47]$$

El problema puede automatizarse con Excel como sigue:

1.7. Supongamos que las calificaciones de tres jueces deportivos sobre el ejercicio de un gimnasta han sido X={1, 2, 3}. Usando probabilidades iguales se extraen muestras aleatorias de dos calificaciones y se consideran los estimadores por analogía media muestral y varianza muestral. Hallar la distribución en el muestreo y sus errores para los dos estimadores en los casos siguientes:

1) Muestreo sin reposición sin tener en cuenta el orden de colocación de los elementos.
2) Muestreo sin reposición teniendo en cuenta el orden de colocación de los elementos.
3) Muestreo con reposición sin tener en cuenta el orden de colocación de los elementos.
4) Muestreo con reposición teniendo en cuenta el orden de colocación de los elementos.

Para ***muestreo sin reposición sin tener en cuenta el orden de colocación de los elementos*** el número de muestras de tamaño 2 en el espacio muestral serán las combinaciones sin repetición de tres elementos tomados de dos en dos:

$$C_{3,2} = \binom{3}{2} = 3$$

Al tratarse de muestreo aleatorio sin reposición y probabilidades iguales, las probabilidades iniciales de selección de los elementos de la población para la muestra valdrán $P(u_i) = 1/3$, i = 1, ..., 3 y la probabilidad de cualquier muestra puede hallarse mediante:

$$P_X = P(u_i,u_j) = P(u_i)P(u_j/u_i)+P(u_j)P(u_i/u_j)=P(u_i)P(u_j)/(1-P(u_i))+P(u_j)P(u_i)/(1-P(u_j))) =$$

$$(1/3^2)/(1-1/3) + (1/3^2)/(1-1/3) = 2(1/3^2)/(1-1/3) = 1/3 = 0{,}33333$$

Se observa que las probabilidades de las muestras serán todas iguales a 1/3. Luego estamos ante un método de selección con probabilidades iguales y muestras equiprobables. El espacio muestral, las probabilidades asociadas a las muestras y la distribución en el muestreo de los estimadores media muestral (*MEDIAM*) y varianza muestral (*VARIANZAM*) se presentan en la siguiente tabla:

S1_X	S2_X	P_X	MEDIAM	VARIANZAM
1	2	1/3	1,5	0,25
1	3	1/3	2	1
2	3	1/3	2,5	0,25

Para comprobar la insesgadez, hallamos la esperanza matemática de los estimadores tal y como se indica a continuación:

$$E(MEDIAM) = \sum_{i=1}^{3} MEDIAM_i P_i = 2 = MEDIAP$$

$$E(VARIANZAM) = \sum_{i=1}^{3} VARIANZA_i P_i = 0{,}5 \neq 2/3 = 0{,}6666 = VARIANZAP$$

Para calcular los sesgos se observa que *MEDIAM* es insesgado para *MEDIAP* y $B(VARIANZAM) = 0{,}5 - 0{,}6666 = -0{,}16666$. A continuación se calculan las varianzas de los estimadores.

$$V(MEDIAM) = \sum_{i=1}^{3} \left(MEDIAM_i - E(MEDIAM)\right)^2 P_i = 0{,}16666$$

$$V(VARIANZAM) = \sum_{i=1}^{3} \left(VARIANZAM_i - E(VARIANZAM)\right)^2 P_i = 0{,}125$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(MEDIAM) = \sqrt{0{,}1666} = 0{,}408, \quad \sigma(VARIANZAM) = \sqrt{0{,}000399} = 0{,}353$$

Como $|B(VARIANZAM)/\sigma(VARIANZAM)| = 0{,}47 > 1/10$, el sesgo del estimador *VARIANZAM* no es despreciable.

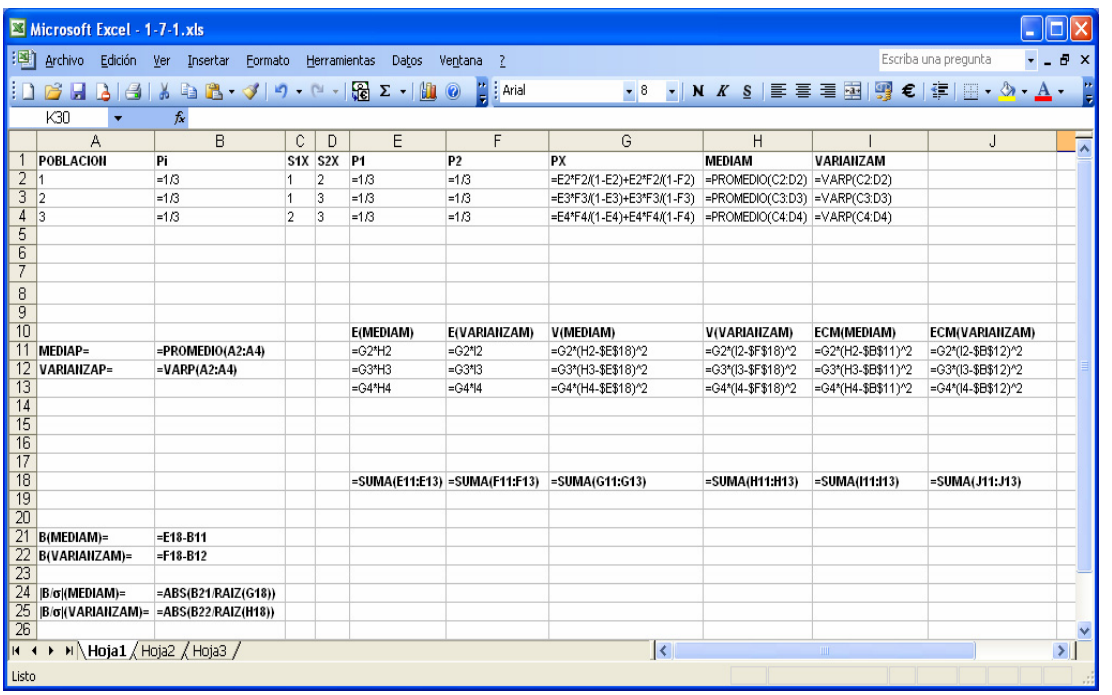
Para hallar el error de muestreo de MEDIAM y VARIANZAM vemos que el segundo estimador es sesgado con sesgo no despreciable y el primero es insesgado. La medición del error debe hacerse a través de los errores cuadráticos medios. Tenemos:

$$ECM(VARIANZAM) = \sum_{i=1}^{3} \left(VARIANZAM_i - 2/3\right)^2 P_i = 0{,}152777$$

$$ECM(MEDIAM) = V(MEDIAM) = 0{,}16666$$

Como el estimador *MEDIAM* es insesgado, su varianza coincide con su error cuadrático medio, luego su precisión se mide a través de la varianza. De esta forma, el estimador *VARIANZAM* para estimar la varianza poblacional es más preciso que el estimador *MEDIAM* para estimar la media poblacional por tener menor error cuadrático medio. Se observa que la ganancia en precisión es pequeña: (0,16666/0,15277-1)100 = 9%.

A continuación se presenta la solución con Excel.



Microsoft Excel - 1-7-1.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

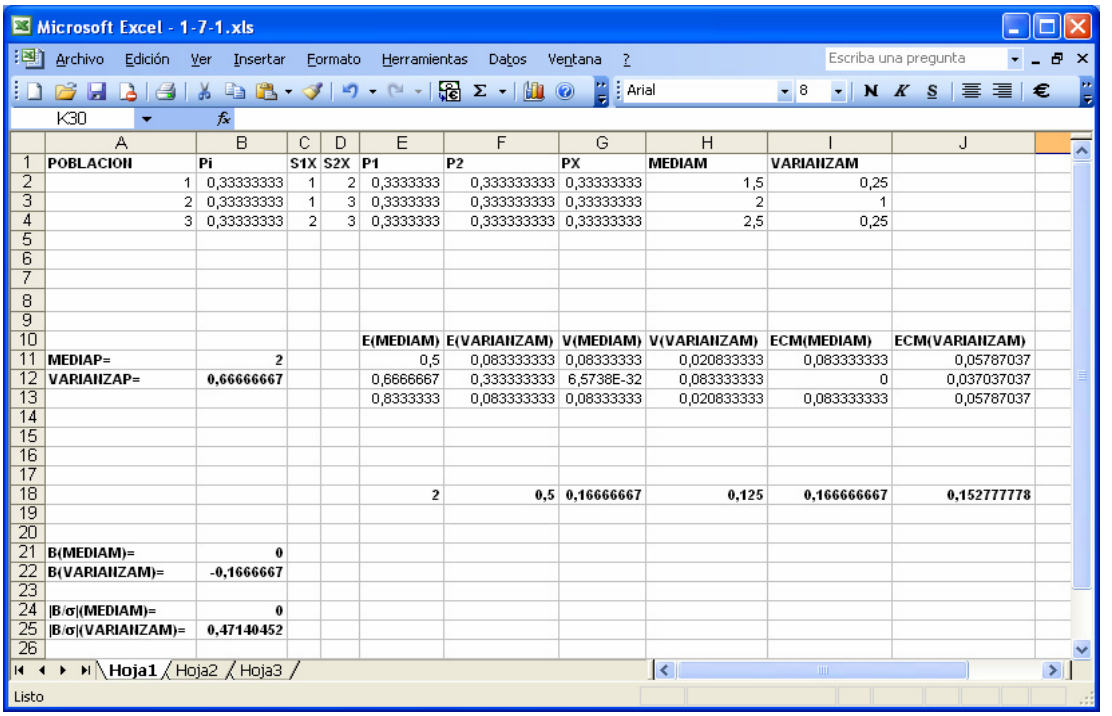
Arial 8

K30

	A	B	C	D	E	F	G	H	I	J
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM	
2	1	=1/3	1	2	=1/3	=1/3	=E2*F2/(1-E2)+E2*F2/(1-F2)	=PROMEDIO(C2:D2)	=VARP(C2:D2)	
3	2	=1/3	1	3	=1/3	=1/3	=E3*F3/(1-E3)+E3*F3/(1-F3)	=PROMEDIO(C3:D3)	=VARP(C3:D3)	
4	3	=1/3	2	3	=1/3	=1/3	=E4*F4/(1-E4)+E4*F4/(1-F4)	=PROMEDIO(C4:D4)	=VARP(C4:D4)	
5										
6										
7										
8										
9										
10					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)	ECM(VARIANZAM)
11	MEDIAP=	=PROMEDIO(A2:A4)			=G2*H2	=G2*I2	=G2*(H2-$E$18)^2	=G2*(I2-$F$18)^2	=G2*(H2-$B$11)^2	=G2*(I2-$B$12)^2
12	VARIANZAP=	=VARP(A2:A4)			=G3*H3	=G3*I3	=G3*(H3-$E$18)^2	=G3*(I3-$F$18)^2	=G3*(H3-$B$11)^2	=G3*(I3-$B$12)^2
13					=G4*H4	=G4*I4	=G4*(H4-$E$18)^2	=G4*(I4-$F$18)^2	=G4*(H4-$B$11)^2	=G4*(I4-$B$12)^2
14										
15										
16										
17										
18					=SUMA(E11:E13)	=SUMA(F11:F13)	=SUMA(G11:G13)	=SUMA(H11:H13)	=SUMA(I11:I13)	=SUMA(J11:J13)
19										
20										
21	B(MEDIAM)=	=E18-B11								
22	B(VARIANZAM)=	=F18-B12								
23										
24	|B/σ|(MEDIAM)=	=ABS(B21/RAIZ(G18))								
25	|B/σ|(VARIANZAM)=	=ABS(B22/RAIZ(H18))								
26										

Hoja1 Hoja2 Hoja3

Listo



Microsoft Excel - 1-7-1.xls

	A	B	C	D	E	F	G	H	I	J
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM	
2	1	0,33333333	1	2	0,3333333	0,333333333	0,33333333	1,5	0,25	
3	2	0,33333333	1	3	0,3333333	0,333333333	0,33333333	2	1	
4	3	0,33333333	2	3	0,3333333	0,333333333	0,33333333	2,5	0,25	
5										
6										
7										
8										
9										
10					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)	ECM(VARIANZAM)
11	MEDIAP=	2			0,5	0,083333333	0,08333333	0,020833333	0,083333333	0,05787037
12	VARIANZAP=	0,66666667			0,6666667	0,333333333	6,5738E-32	0,083333333	0	0,037037037
13					0,8333333	0,083333333	0,08333333	0,020833333	0,083333333	0,05787037
14										
15										
16										
17										
18					2	0,5	0,16666667	0,125	0,166666667	0,152777778
19										
20										
21	B(MEDIAM)=	0								
22	B(VARIANZAM)=	-0,1666667								
23										
24	|B/σ|(MEDIAM)=	0								
25	|B/σ|(VARIANZAM)=	0,47140452								
26										

Hoja1 / Hoja2 / Hoja3

Listo

Para ***muestreo sin reposición teniendo en cuenta el orden de colocación de los elementos*** el número de muestras de tamaño 2 en el espacio muestral serán las variaciones sin repetición de tres elementos tomados de dos en dos:

$$V_{3,2} = \binom{3}{2} \cdot 2! = 6$$

Al tratarse de muestreo aleatorio sin reposición y probabilidades iguales, las probabilidades iniciales de selección de los elementos de la población para la muestra valdrán $P(u_i) = 1/3$, $i = 1, ..., 3$ y la probabilidad de cualquier muestra puede hallarse mediante:

$$P_X = P(u_i,u_j) = P(u_i)P(u_j/u_i) = P(u_i)P(u_j)/(1-P(u_i)) = (1/3^2)/(1-1/3) = 1/6 = 0,1666$$

Se observa que las probabilidades de las muestras serán todas iguales a 1/6. Luego estamos ante un método de selección con probabilidades iguales y muestras equiprobables. El espacio muestral, las probabilidades asociadas a las muestras y la distribución en el muestreo de los estimadores media muestral (*MEDIAM*) y varianza muestral (*VARIANZAM*) se presentan en la siguiente tabla:

S1_X	S2_X	P_X	MEDIAM	VARIANZAM
1	2	1/6	1,5	0,25
1	3	1/6	2	1
2	1	1/6	1,5	0,25
2	3	1/6	2,5	0,25
3	1	1/6	2	1
3	2	1/6	2,5	0,25

Para comprobar la insesgadez, hallamos la esperanza matemática de los estimadores tal y como se indica a continuación:

$$E(MEDIAM) = \sum_{i=1}^{6} MEDIAM_i P_i = 2 = MEDIAP$$

$$E(VARIANZAM) = \sum_{i=1}^{6} VARIANZA_i P_i = 0{,}5 \neq 2/3 = 0{,}6666 = VARIANZAP$$

Para calcular los sesgos se observa que *MEDIAM* es insesgado para *MEDIAP* y $B(VARIANZAM) = 0{,}5 - 0{,}6666 = -0{,}16666$. A continuación se calculan las varianzas de los estimadores.

$$V(MEDIAM) = \sum_{i=1}^{6} \left(MEDIAM_i - E(MEDIAM)\right)^2 P_i = 0{,}16666$$

$$V(VARIANZAM) = \sum_{i=1}^{6} \left(VARIANZAM_i - E(VARIANZAM)\right)^2 P_i = 0{,}125$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(MEDIAM) = \sqrt{0{,}1666} = 0{,}408, \quad \sigma(VARIANZAM) = \sqrt{0{,}000399} = 0{,}353$$

Como $|B(VARIANZAM)/\sigma(VARIANZAM)| = 0{,}47 > 1/10$, el sesgo del estimador VARIANZAM no es despreciable.

Para hallar el error de muestreo de *MEDIAM* y *VARIANZAM* vemos que el segundo estimador es sesgado con sesgo no despreciable y el primero es insesgado. La medición del error debe hacerse a través de los errores cuadráticos medios. Tenemos:

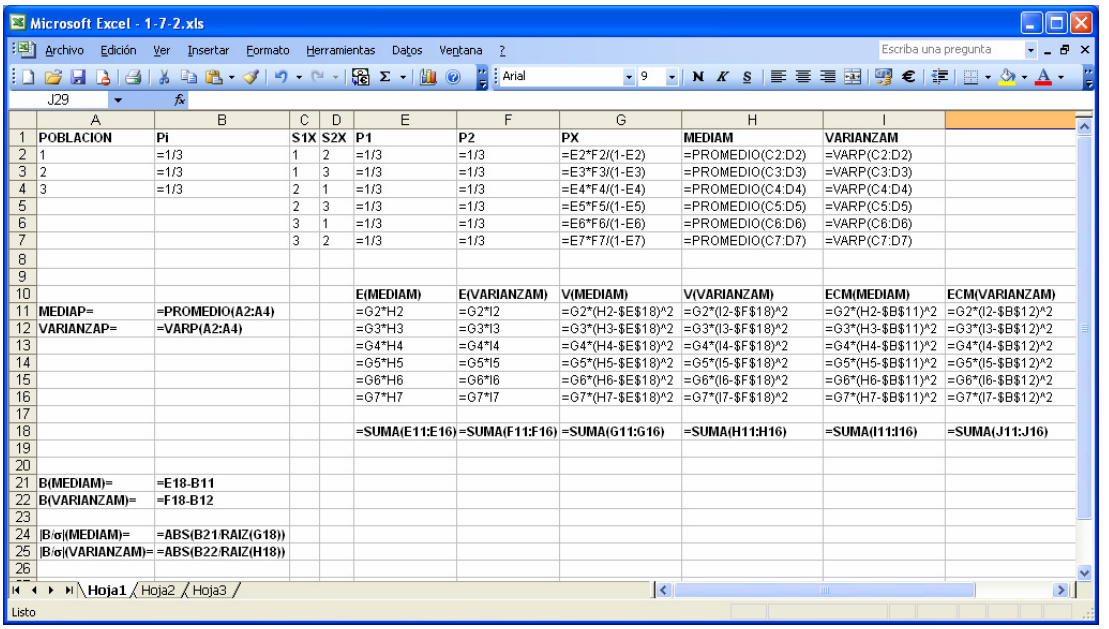
$$ECM(VARIANZAM) = \sum_{i=1}^{6} \left(VARIANZAM_i - 2/3\right)^2 P_i = 0{,}152777$$

$$ECM(MEDIAM) = V(MEDIAM) = 0{,}16666$$

Como el estimador MEDIAM es insesgado, su varianza coincide con su error cuadrático medio, luego su precisión se mide a través de la varianza. De esta forma, el estimador VARIANZAM para estimar la varianza poblacional es más preciso que el estimador MEDIAM para estimar la media poblacional por tener menor error cuadrático medio. Se observa que la ganancia en precisión es pequeña: (0,16666/0,15277-1)100 = 9%.

Se observa que cuando se trabaja sin reposición, el hecho de tener o no en cuenta el orden de colocación de los elementos en las muestras no interviene a los efectos de cálculo de medias, varianzas, sesgos, intervalos de confianza y precisiones de los estimadores.

Los cálculos pueden implementarse mediante Excel tal y como se indica en las pantallas siguientes:



Microsoft Excel - 1-7-2.xls

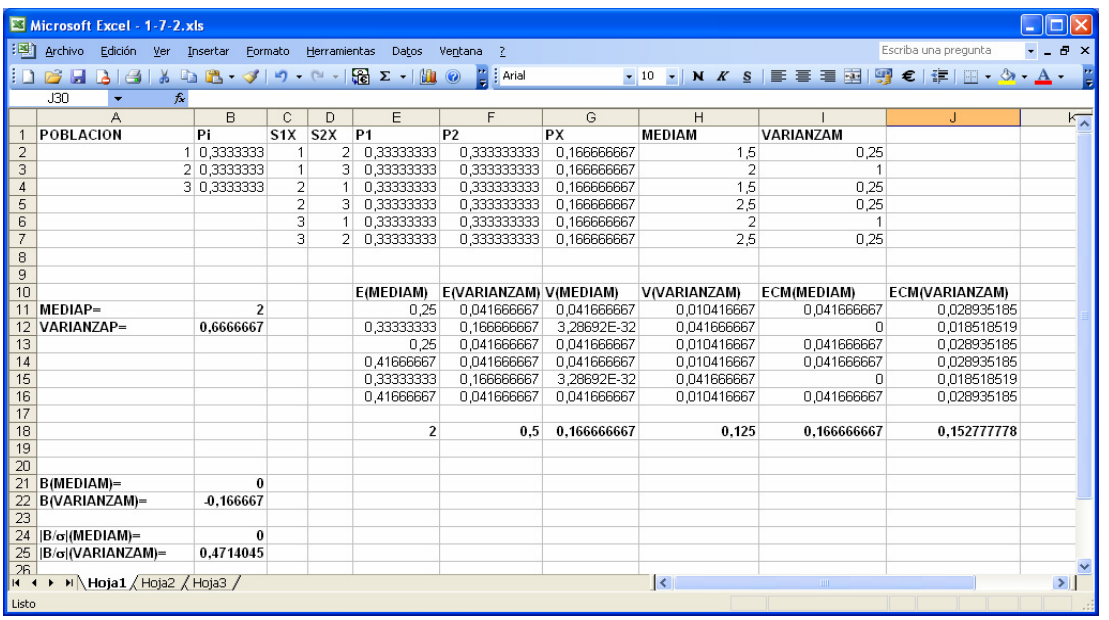
Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

J29

	A	B	C	D	E	F	G	H	I
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM
2	1	=1/3	1	2	=1/3	=1/3	=E2*F2/(1-E2)	=PROMEDIO(C2:D2)	=VARP(C2:D2)
3	2	=1/3	1	3	=1/3	=1/3	=E3*F3/(1-E3)	=PROMEDIO(C3:D3)	=VARP(C3:D3)
4	3	=1/3	2	1	=1/3	=1/3	=E4*F4/(1-E4)	=PROMEDIO(C4:D4)	=VARP(C4:D4)
5			2	3	=1/3	=1/3	=E5*F5/(1-E5)	=PROMEDIO(C5:D5)	=VARP(C5:D5)
6			3	1	=1/3	=1/3	=E6*F6/(1-E6)	=PROMEDIO(C6:D6)	=VARP(C6:D6)
7			3	2	=1/3	=1/3	=E7*F7/(1-E7)	=PROMEDIO(C7:D7)	=VARP(C7:D7)
8									
9									
10					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)	ECM(VARIANZAM)
11	MEDIAP=	=PROMEDIO(A2:A4)			=G2*H2	=G2*I2	=G2*(H2-$E$18)^2	=G2*(I2-$F$18)^2	=G2*(H2-$B$11)^2	=G2*(I2-$B$12)^2
12	VARIANZAP=	=VARP(A2:A4)			=G3*H3	=G3*I3	=G3*(H3-$E$18)^2	=G3*(I3-$F$18)^2	=G3*(H3-$B$11)^2	=G3*(I3-$B$12)^2
13					=G4*H4	=G4*I4	=G4*(H4-$E$18)^2	=G4*(I4-$F$18)^2	=G4*(H4-$B$11)^2	=G4*(I4-$B$12)^2
14					=G5*H5	=G5*I5	=G5*(H5-$E$18)^2	=G5*(I5-$F$18)^2	=G5*(H5-$B$11)^2	=G5*(I5-$B$12)^2
15					=G6*H6	=G6*I6	=G6*(H6-$E$18)^2	=G6*(I6-$F$18)^2	=G6*(H6-$B$11)^2	=G6*(I6-$B$12)^2
16					=G7*H7	=G7*I7	=G7*(H7-$E$18)^2	=G7*(I7-$F$18)^2	=G7*(H7-$B$11)^2	=G7*(I7-$B$12)^2
17									
18					=SUMA(E11:E16)	=SUMA(F11:F16)	=SUMA(G11:G16)	=SUMA(H11:H16)	=SUMA(I11:I16)	=SUMA(J11:J16)
19									
20									
21	B(MEDIAM)=	=E18-B11								
22	B(VARIANZAM)=	=F18-B12								
23										
24	|B/σ|(MEDIAM)=	=ABS(B21/RAIZ(G18))								
25	|B/σ|(VARIANZAM)=	=ABS(B22/RAIZ(H18))								
26										

Hoja1 Hoja2 Hoja3

Listo



Microsoft Excel - 1-7-2.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

J30

	A	B	C	D	E	F	G	H	I	J
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM	
2	1	0,3333333	1	2	0,33333333	0,333333333	0,166666667	1,5	0,25	
3	2	0,3333333	1	3	0,33333333	0,333333333	0,166666667	2	1	
4	3	0,3333333	2	1	0,33333333	0,333333333	0,166666667	1,5	0,25	
5			2	3	0,33333333	0,333333333	0,166666667	2,5	0,25	
6			3	1	0,33333333	0,333333333	0,166666667	2	1	
7			3	2	0,33333333	0,333333333	0,166666667	2,5	0,25	
8										
9										
10					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)	ECM(VARIANZAM)
11	MEDIAP=	2			0,25	0,041666667	0,041666667	0,010416667	0,041666667	0,028935185
12	VARIANZAP=	0,6666667			0,33333333	0,166666667	3,28692E-32	0,041666667	0	0,018518519
13					0,25	0,041666667	0,041666667	0,010416667	0,041666667	0,028935185
14					0,41666667	0,041666667	0,041666667	0,010416667	0,041666667	0,028935185
15					0,33333333	0,166666667	3,28692E-32	0,041666667	0	0,018518519
16					0,41666667	0,041666667	0,041666667	0,010416667	0,041666667	0,028935185
17										
18					2	0,5	0,166666667	0,125	0,166666667	0,152777778
19										
20										
21	B(MEDIAM)=	0								
22	B(VARIANZAM)=	-0,166667								
23										
24	|B/σ|(MEDIAM)=	0								
25	|B/σ|(VARIANZAM)=	0,4714045								
26										

Hoja1 Hoja2 Hoja3

Listo

Para ***muestreo con reposición sin tener en cuenta el orden de colocación de los elementos*** el número de muestras de tamaño dos en el espacio muestral serán las combinaciones con repetición de tres elementos tomados de dos en dos:

$$CR_{3,2} = \binom{3+2-1}{2} = 6$$

Al tratarse de muestreo aleatorio con reposición y probabilidades iguales, las probabilidades iniciales de selección de los elementos de la población para la muestra valdrán $P(u_i) = 1/3$, i = 1, ..., 3 y la probabilidad de cualquier muestra puede hallarse mediante:

$$P_X = P(u_i,u_j) = 2\,P(u_i)P(u_j) = 2(1/3)^2 = 2/9 \text{ si } i \neq j$$
$$P_X = P(u_i,u_i) = P(u_i)^2 = (1/3)^2 = 1/9 \text{ si } i = j$$

Se observa que las probabilidades de las muestras serán todas iguales a 1/3. Luego estamos ante un método de selección con probabilidades iguales y muestras equiprobables. El espacio muestral, las probabilidades asociadas a las muestras y la distribución en el muestreo de los estimadores media muestral (*MEDIAM*) y varianza muestral (*VARIANZAM*) se presentan en la siguiente tabla:

S1_X	S2_X	P_X	MEDIAM	VARIANZAM
1	1	1/9	1	0
1	2	2/9	1,5	0,25
1	3	2/9	2	1
2	2	1/9	2	0
2	3	2/9	2,5	0,25
3	3	1/9	3	0

Para comprobar la insesgadez, hallamos la esperanza matemática de los estimadores tal y como se indica a continuación:

$$E(MEDIAM) = \sum_{i=1}^{6} MEDIAM_i P_i = 2 = MEDIAP$$

$$E(VARIANZAM) = \sum_{i=1}^{6} VARIANZA_i P_i = 1/3 = 0{,}3333 \neq 2/3 = 0{,}6666 = VARIANZAP$$

Para calcular los sesgos se observa que *MEDIAM* es insesgado para *MEDIAP* y $B(VARIANZAM) = 1/3\text{-}2/3 = -1/3 = -0{,}3333$. A continuación se calculan las varianzas de los estimadores.

$$V(MEDIAM) = \sum_{i=1}^{6} \left(MEDIAM_i - E(MEDIAM)\right)^2 P_i = 0{,}3333$$

$$V(VARIANZAM) = \sum_{i=1}^{6} \left(VARIANZAM_i - E(VARIANZAM)\right)^2 P_i = 0{,}13888$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(MEDIAM) = \sqrt{0{,}3333} = 0{,}577, \quad \sigma(VARIANZAM) = \sqrt{0{,}13888} = 0{,}372$$

Como $|B(VARIANZAM)/\sigma(VARIANZAM)| = 0{,}894 > 1/10$ el sesgo del estimador VARIANZAM no es despreciable.

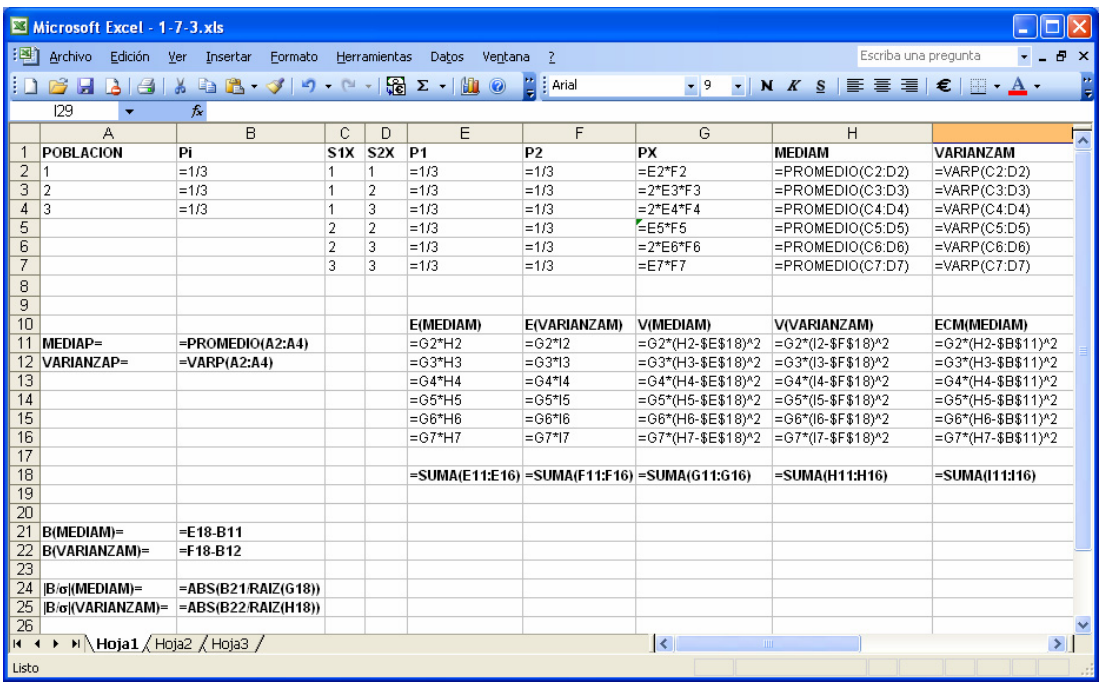
Para hallar el error de muestreo de MEDIAM y VARIANZAM vemos que el segundo estimador es sesgado con sesgo no despreciable y el primero es insesgado. La medición del error debe hacerse a través de los errores cuadráticos medios. Tenemos:

$$ECM(VARIANZAM) = \sum_{i=1}^{6} \left(VARIANZAM_i - 2/3\right)^2 P_i = 0{,}25$$

$$ECM(MEDIAM) = V(MEDIAM) = 0{,}3333$$

Como el estimador *MEDIAM* es insesgado, su varianza coincide con su error cuadrático medio, luego su precisión se mide a través de la varianza. De esta forma, el estimador VARIANZAM para estimar la varianza poblacional es más preciso que el estimador *MEDIAM* para estimar la media poblacional por tener menor error cuadrático medio. Se observa que la ganancia en precisión es (0,3333/0,25-1)*100 = 33,32%. Hay que subrayar que la ganancia en precisión es ahora mayor que en el mismo caso para muestreo sin reposición, debido a la mayor precisión en general del muestreo sin reposición.

Los cálculos en Excel se muestran a continuación.



Microsoft Excel - 1-7-3.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

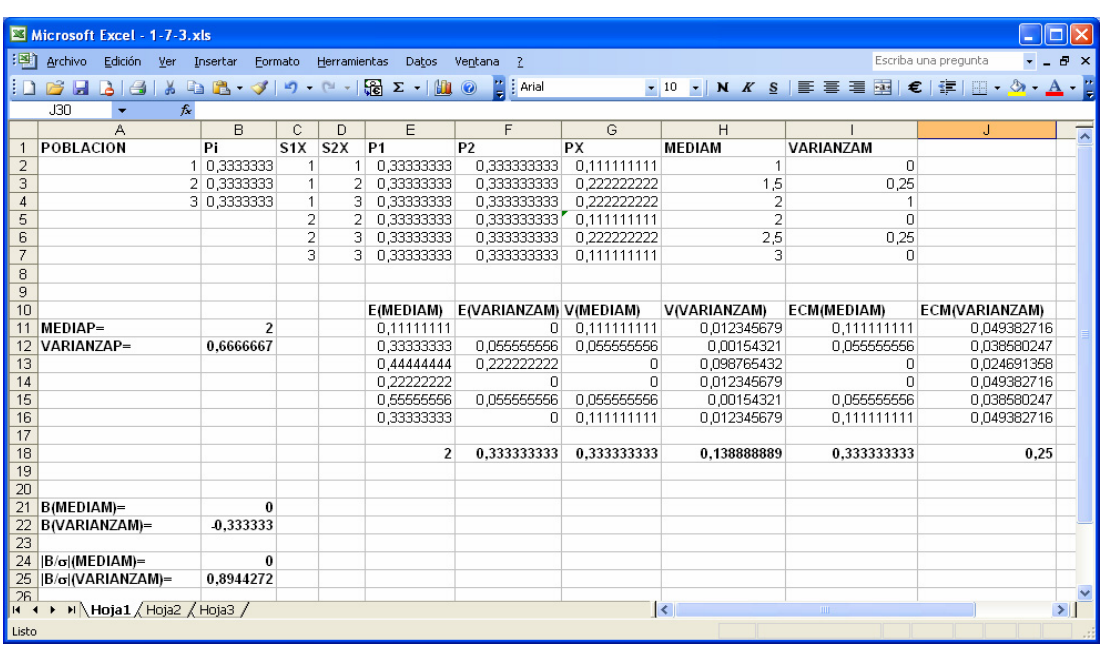
Arial 9

I29

	A	B	C	D	E	F	G	H	
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM
2	1	=1/3	1	1	=1/3	=1/3	=E2*F2	=PROMEDIO(C2:D2)	=VARP(C2:D2)
3	2	=1/3	1	2	=1/3	=1/3	=2*E3*F3	=PROMEDIO(C3:D3)	=VARP(C3:D3)
4	3	=1/3	1	3	=1/3	=1/3	=2*E4*F4	=PROMEDIO(C4:D4)	=VARP(C4:D4)
5			2	2	=1/3	=1/3	=E5*F5	=PROMEDIO(C5:D5)	=VARP(C5:D5)
6			2	3	=1/3	=1/3	=2*E6*F6	=PROMEDIO(C6:D6)	=VARP(C6:D6)
7			3	3	=1/3	=1/3	=E7*F7	=PROMEDIO(C7:D7)	=VARP(C7:D7)
8									
9									
10					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)
11	MEDIAP=	=PROMEDIO(A2:A4)			=G2*H2	=G2*I2	=G2*(H2-$E$18)^2	=G2*(I2-$F$18)^2	=G2*(H2-$B$11)^2
12	VARIANZAP=	=VARP(A2:A4)			=G3*H3	=G3*I3	=G3*(H3-$E$18)^2	=G3*(I3-$F$18)^2	=G3*(H3-$B$11)^2
13					=G4*H4	=G4*I4	=G4*(H4-$E$18)^2	=G4*(I4-$F$18)^2	=G4*(H4-$B$11)^2
14					=G5*H5	=G5*I5	=G5*(H5-$E$18)^2	=G5*(I5-$F$18)^2	=G5*(H5-$B$11)^2
15					=G6*H6	=G6*I6	=G6*(H6-$E$18)^2	=G6*(I6-$F$18)^2	=G6*(H6-$B$11)^2
16					=G7*H7	=G7*I7	=G7*(H7-$E$18)^2	=G7*(I7-$F$18)^2	=G7*(H7-$B$11)^2
17									
18					=SUMA(E11:E16)	=SUMA(F11:F16)	=SUMA(G11:G16)	=SUMA(H11:H16)	=SUMA(I11:I16)
19									
20									
21	B(MEDIAM)=	=E18-B11							
22	B(VARIANZAM)=	=F18-B12							
23									
24	|B/σ|(MEDIAM)=	=ABS(B21/RAIZ(G18))							
25	|B/σ|(VARIANZAM)=	=ABS(B22/RAIZ(H18))							
26									

Hoja1 Hoja2 Hoja3

Listo



Microsoft Excel - 1-7-3.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

Arial 10

J30

	A	B	C	D	E	F	G	H	I	J
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM	
2	1	0,3333333	1	1	0,33333333	0,333333333	0,111111111	1	0	
3	2	0,3333333	1	2	0,33333333	0,333333333	0,222222222	1,5	0,25	
4	3	0,3333333	1	3	0,33333333	0,333333333	0,222222222	2	1	
5			2	2	0,33333333	0,333333333	0,111111111	2	0	
6			2	3	0,33333333	0,333333333	0,222222222	2,5	0,25	
7			3	3	0,33333333	0,333333333	0,111111111	3	0	
8										
9										
10					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)	ECM(VARIANZAM)
11	MEDIAP=	2			0,11111111	0	0,111111111	0,012345679	0,111111111	0,049382716
12	VARIANZAP=	0,6666667			0,33333333	0,055555556	0,055555556	0,00154321	0,055555556	0,038580247
13					0,44444444	0,222222222	0	0,098765432	0	0,024691358
14					0,22222222	0	0	0,012345679	0	0,049382716
15					0,55555556	0,055555556	0,055555556	0,00154321	0,055555556	0,038580247
16					0,33333333	0	0,111111111	0,012345679	0,111111111	0,049382716
17										
18					2	0,333333333	0,333333333	0,138888889	0,333333333	0,25
19										
20										
21	B(MEDIAM)=	0								
22	B(VARIANZAM)=	-0,333333								
23										
24	|B/σ|(MEDIAM)=	0								
25	|B/σ|(VARIANZAM)=	0,8944272								
26										

Hoja1 Hoja2 Hoja3

Listo

Para ***muestreo con reposición teniendo en cuenta el orden de colocación de los elementos*** el número de muestras de tamaño 2 en el espacio muestral serán las variaciones con repetición de tres elementos tomados de dos en dos:

$$VR_{3,2} = 3^2 = 9$$

Al tratarse de muestreo aleatorio con reposición y probabilidades iguales, las probabilidades iniciales de selección de los elementos de la población para la muestra valdrán $P(u_i) = 1/3$, $i = 1, ..., 3$ y la probabilidad de cualquier muestra puede hallarse mediante:

$$P_X = P(u_i,u_j) = P(u_i)P(u_j) = (1/3)^2 = 1/9$$

Se observa que las probabilidades de las muestras serán todas iguales a 1/3. Luego estamos ante un método de selección con probabilidades iguales y muestras equiprobables. El espacio muestral, las probabilidades asociadas a las muestras y la distribución en el muestreo de los estimadores media muestral (*MEDIAM*) y varianza muestral (*VARIANZAM*) se presentan en la siguiente tabla:

S1_X	S2_X	P_X	MEDIAM	VARIANZAM
1	1	1/9	1	0
1	2	1/9	1,5	0,25
1	3	1/9	2	1
2	1	1/9	1,5	0,25
2	2	1/9	2	0
2	3	1/9	2,5	0,25
3	1	1/9	2	1
3	2	1/9	2,5	0,25
3	3	1/9	3	0

Para comprobar la insesgadez, hallamos la esperanza matemática de los estimadores tal y como se indica a continuación:

$$E(MEDIAM) = \sum_{i=1}^{9} MEDIAM_i P_i = 2 = MEDIAP$$

$$E(VARIANZAM) = \sum_{i=1}^{9} VARIANZA_i P_i = 1/3 = 0{,}3333 \neq 2/3 = 0{,}6666 = VARIANZAP$$

Para calcular los sesgos se observa que *MEDIAM* es insesgado para *MEDIAP* y $B(VARIANZAM) = 1/3 - 2/3 = -1/3 = -0{,}3333$. A continuación se calculan las varianzas de los estimadores.

$$V(MEDIAM) = \sum_{i=1}^{9} \left(MEDIAM_i - E(MEDIAM)\right)^2 P_i = 0{,}3333$$

$$V(VARIANZAM) = \sum_{i=1}^{9} \left(VARIANZAM_i - E(VARIANZAM)\right)^2 P_i = 0{,}13888$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(MEDIAM) = \sqrt{0{,}3333} = 0{,}577, \quad \sigma(VARIANZAM) = \sqrt{0{,}13888} = 0{,}372$$

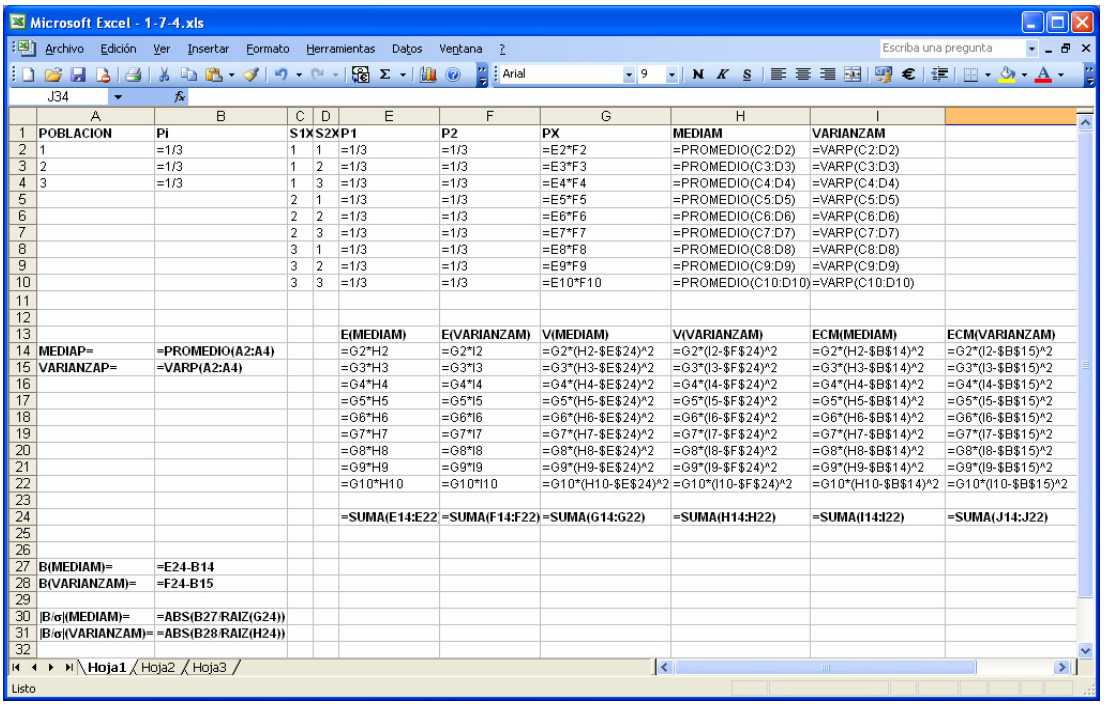
Como $|B(VARIANZAM)/\sigma(VARIANZAM)| = 0{,}894 > 1/10$, el sesgo del estimador VARIANZAM no es despreciable.

Para hallar el error de muestreo de *MEDIAM* y *VARIANZAM* vemos que el segundo estimador es sesgado con sesgo no despreciable y el primero es insesgado. La medición del error debe hacerse a través de los errores cuadráticos medios. Tenemos:

$$ECM(VARIANZAM) = \sum_{i=1}^{9} \left(VARIANZAM_i - 2/3\right)^2 P_i = 0{,}25$$

$$ECM(MEDIAM) = V(MEDIAM) = 0{,}3333$$

Los cálculos se automatizan con Excel como sigue:

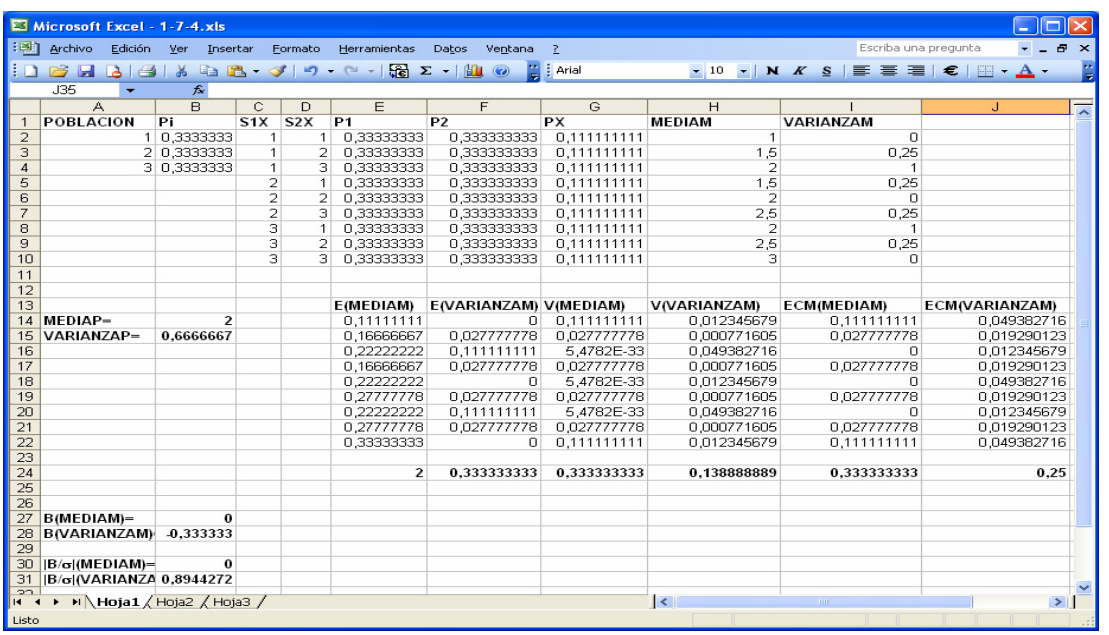


Microsoft Excel - 1-7-4.xls

	A	B	C	D	E	F	G	H	I
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM
2	1	=1/3	1	1	=1/3	=1/3	=E2*F2	=PROMEDIO(C2:D2)	=VARP(C2:D2)
3	2	=1/3	1	2	=1/3	=1/3	=E3*F3	=PROMEDIO(C3:D3)	=VARP(C3:D3)
4	3	=1/3	1	3	=1/3	=1/3	=E4*F4	=PROMEDIO(C4:D4)	=VARP(C4:D4)
5			2	1	=1/3	=1/3	=E5*F5	=PROMEDIO(C5:D5)	=VARP(C5:D5)
6			2	2	=1/3	=1/3	=E6*F6	=PROMEDIO(C6:D6)	=VARP(C6:D6)
7			2	3	=1/3	=1/3	=E7*F7	=PROMEDIO(C7:D7)	=VARP(C7:D7)
8			3	1	=1/3	=1/3	=E8*F8	=PROMEDIO(C8:D8)	=VARP(C8:D8)
9			3	2	=1/3	=1/3	=E9*F9	=PROMEDIO(C9:D9)	=VARP(C9:D9)
10			3	3	=1/3	=1/3	=E10*F10	=PROMEDIO(C10:D10)	=VARP(C10:D10)
11									
12									
13					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)	ECM(VARIANZAM)
14	MEDIAP=	=PROMEDIO(A2:A4)			=G2*H2	=G2*I2	=G2*(H2-$E$24)^2	=G2*(I2-$F$24)^2	=G2*(H2-$B$14)^2	=G2*(I2-$B$15)^2
15	VARIANZAP=	=VARP(A2:A4)			=G3*H3	=G3*I3	=G3*(H3-$E$24)^2	=G3*(I3-$F$24)^2	=G3*(H3-$B$14)^2	=G3*(I3-$B$15)^2
16					=G4*H4	=G4*I4	=G4*(H4-$E$24)^2	=G4*(I4-$F$24)^2	=G4*(H4-$B$14)^2	=G4*(I4-$B$15)^2
17					=G5*H5	=G5*I5	=G5*(H5-$E$24)^2	=G5*(I5-$F$24)^2	=G5*(H5-$B$14)^2	=G5*(I5-$B$15)^2
18					=G6*H6	=G6*I6	=G6*(H6-$E$24)^2	=G6*(I6-$F$24)^2	=G6*(H6-$B$14)^2	=G6*(I6-$B$15)^2
19					=G7*H7	=G7*I7	=G7*(H7-$E$24)^2	=G7*(I7-$F$24)^2	=G7*(H7-$B$14)^2	=G7*(I7-$B$15)^2
20					=G8*H8	=G8*I8	=G8*(H8-$E$24)^2	=G8*(I8-$F$24)^2	=G8*(H8-$B$14)^2	=G8*(I8-$B$15)^2
21					=G9*H9	=G9*I9	=G9*(H9-$E$24)^2	=G9*(I9-$F$24)^2	=G9*(H9-$B$14)^2	=G9*(I9-$B$15)^2
22					=G10*H10	=G10*I10	=G10*(H10-$E$24)^2	=G10*(I10-$F$24)^2	=G10*(H10-$B$14)^2	=G10*(I10-$B$15)^2
23										
24					=SUMA(E14:E22	=SUMA(F14:F22)	=SUMA(G14:G22)	=SUMA(H14:H22)	=SUMA(I14:I22)	=SUMA(J14:J22)
25										
26										
27	B(MEDIAM)=	=E24-B14								
28	B(VARIANZAM)=	=F24-B15								
29										
30	|B/σ|(MEDIAM)=	=ABS(B27/RAIZ(G24))								
31	|B/σ|(VARIANZAM)=	=ABS(B28/RAIZ(H24))								
32										

Hoja1 / Hoja2 / Hoja3

Listo



Microsoft Excel - 1-7-4.xls

	A	B	C	D	E	F	G	H	I	J
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM	
2	1	0,3333333	1	1	0,33333333	0,33333333	0,11111111	1	0	
3	2	0,3333333	1	2	0,33333333	0,33333333	0,11111111	1,5	0,25	
4	3	0,3333333	1	3	0,33333333	0,33333333	0,11111111	2	1	
5			2	1	0,33333333	0,33333333	0,11111111	1,5	0,25	
6			2	2	0,33333333	0,33333333	0,11111111	2	0	
7			2	3	0,33333333	0,33333333	0,11111111	2,5	0,25	
8			3	1	0,33333333	0,33333333	0,11111111	2	1	
9			3	2	0,33333333	0,33333333	0,11111111	2,5	0,25	
10			3	3	0,33333333	0,33333333	0,11111111	3	0	
11										
12										
13					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)	ECM(VARIANZAM)
14	MEDIAP=	2			0,11111111	0	0,11111111	0,012345679	0,11111111	0,049382716
15	VARIANZAP=	0,6666667			0,16666667	0,027777778	0,027777778	0,000771605	0,027777778	0,019290123
16					0,22222222	0,11111111	5,4782E-33	0,049382716	0	0,012345679
17					0,16666667	0,027777778	0,027777778	0,000771605	0,027777778	0,019290123
18					0,22222222	0	5,4782E-33	0,012345679	0	0,049382716
19					0,27777778	0,027777778	0,027777778	0,000771605	0,027777778	0,019290123
20					0,22222222	0,11111111	5,4782E-33	0,049382716	0	0,012345679
21					0,27777778	0,027777778	0,027777778	0,000771605	0,027777778	0,019290123
22					0,33333333	0	0,11111111	0,012345679	0,11111111	0,049382716
23										
24					2	0,33333333	0,33333333	0,13888889	0,33333333	0,25
25										
26										
27	B(MEDIAM)=	0								
28	B(VARIANZAM)=	-0,333333								
29										
30	|B/σ|(MEDIAM)=	0								
31	|B/σ|(VARIANZA	0,8944272								

Hoja1 / Hoja2 / Hoja3

Listo

Como el estimador *MEDIAM* es insesgado, su varianza coincide con su error cuadrático medio, luego su precisión se mide a través de la varianza. De esta forma, el estimador *VARIANZAM* para estimar la varianza poblacional es más preciso que el estimador *MEDIAM* para estimar la media poblacional por tener menor error cuadrático medio. Se observa que la ganancia en precisión es (0,3333/0,25 – 1)100 = 33,32%. Hay que subrayar que la ganancia en precisión es ahora mayor que en el mismo caso para muestreo sin reposición, lo que es debido a la mayor precisión en general del muestreo sin reposición.

Se observa que cuando se trabaja con reposición, el hecho de tener o no en cuenta el orden de colocación de los elementos en las muestras no interviene a los efectos de cálculo de medias, varianzas, sesgos, intervalos de confianza y precisiones de los estimadores. Ya vimos que esto mismo ocurría cuando se trabajaba sin reposición.

Si comparamos los métodos con reposición con los métodos sin reposición vemos que los errores de muestreo siempre son mayores con reposición. Para el estimador *MEDIAM* la ganancia en precisión por muestrear sin reposición se cuantifica en (0,333/0,166 – 1)100 = 100%, ya que se duplica la precisión. Para el estimador *VARIANZAM* la ganancia en precisión por muestrear sin reposición se cuantifica en (0,25/0,15277 – 1)100 = 63,63%. La ganancia en precisión para *VARIANZAM* es menor que para *MEDIAM,* porque habíamos visto que *VARIANZAM* es más preciso que media *M* y los estimadores más precisos son los que menos precisión pierden al considerar muestreo con reposición.

1.8. En una prueba de patinaje artístico los 10 jueces del jurado calificaron a un patinador con tres cincos, cuatro seises y tres sietes. Usando probabilidades iguales se extraen muestras aleatorias de dos calificaciones sin reposición y teniendo en cuenta el orden de colocación de los elementos. Se consideran los estimadores por analogía media muestral, varianza muestral y recorrido para estimar la calificación media y su dispersión (por dos vías). Hallar la distribución en el muestreo y sus errores para los tres estimadores.

Las probabilidades iniciales de selección serán las siguientes:

X_i	5	6	7
P_i	3/10	4/10	3/10

Para ***muestreo sin reposición teniendo en cuenta el orden de colocación de los elementos*** el número de ***muestras de tamaño 2*** en el espacio muestral serán las variaciones sin repetición de 10 elementos tomados de dos en dos:

$$V_{3,2} = \binom{3}{2} \cdot 2! = 6$$

Al tratarse de muestreo aleatorio sin reposición teniendo en cuenta el orden, la probabilidad de cualquier muestra puede hallarse mediante:

$$P_X = \mathrm{P}(u_i,u_j) = \mathrm{P}(u_i)\mathrm{P}(u_j/u_i) = \mathrm{P}(u_i)\mathrm{P}(u_j)/(1\text{-}\mathrm{P}(u_i)) = P_iP_j/(1\text{-}P_i)$$

El espacio muestral, las probabilidades asociadas a las muestras y la distribución en el muestreo de los estimadores media muestral (*MEDIAM*), varianza muestral (*VARIANZAM*) y recorrido muestral *RM* se presentan en la siguiente tabla:

S1_X	S2_X	P1	P2	P_X	MEDIAM	VARIANZAM	RM
5	6	0,3	0,4	0,171=0,3*0,4/(1-0,3)	5,5	0,25	1
5	7	0,3	0,3	0,128=0,3*0,3/(1-0,3)	6	1	2
6	7	0,4	0,3	0,2=0,4*0,3/(1-0,4)	6,5	0,25	1
6	5	0,4	0,3	0,2=0,4*0,3/(1-0,4)	5,5	0,25	1
7	5	0,3	0,3	0,128=0,3*0,3/(1-0,3)	6	1	2
7	6	0,3	0,4	0,171=0,3*0,4/(1-0,3)	6,5	0,25	1

Para comprobar la insesgadez, hallamos la esperanza matemática de los estimadores tal y como se indica a continuación:

$$E(MEDIAM) = \sum_{i=1}^{6} MEDIAM_i P_i = 6 = MEDIAP$$

$$E(VARIANZAM) = \sum_{i=1}^{6} VARIANZA_i P_i = 0{,}442 \neq 0{,}6 = VARIANZAP$$

$$E(RM) = \sum_{i=1}^{6} RM_i P_i = 1{,}257 \neq 2 = RP$$

Para calcular los sesgos se observa que *MEDIAM* es insesgado para *MEDIAP*, $B(VARIANZAM) = 0{,}442\text{-}0{,}6 = -0{,}157$, y $B(RM) = 1{,}257 - 2 = -0{,}743$. A continuación se calculan las varianzas de los estimadores.

$$V(MEDIAM) = \sum_{i=1}^{6} \left(MEDIAM_i - E(MEDIAM)\right)^2 P_i = 0{,}185$$

$$V(VARIANZAM) = \sum_{i=1}^{6} \left(VARIANZAM_i - E(VARIANZAM)\right)^2 P_i = 0{,}107$$

$$V(RM) = \sum_{i=1}^{6} \left(RM_i - E(RM)\right)^2 P_i = 0{,}191$$

Con lo que las desviaciones típicas valdrán:

$$\sigma(MEDIAM) = \sqrt{0{,}1666} = 0{,}408, \quad \sigma(VARIANZAM) = \sqrt{0{,}000399} = 0{,}353$$

Como $|B(VARIANZAM)/\sigma(VARIANZAM)| = 0{,}47 > 1/10$, el sesgo del estimador VARIANZAM es no despreciable.

Como $|B(RM)/\sigma(RM)| = 1{,}7 > 1/10$, el sesgo del estimador *RM* no es despreciable.

Para hallar el error de muestreo de *MEDIAM*, *VARIANZAM* y *RM* vemos que los dos últimos estimadores son sesgados con sesgo no despreciable y el primero es insesgado. La medición del error debe hacerse a través de los errores cuadráticos medios. Tenemos:

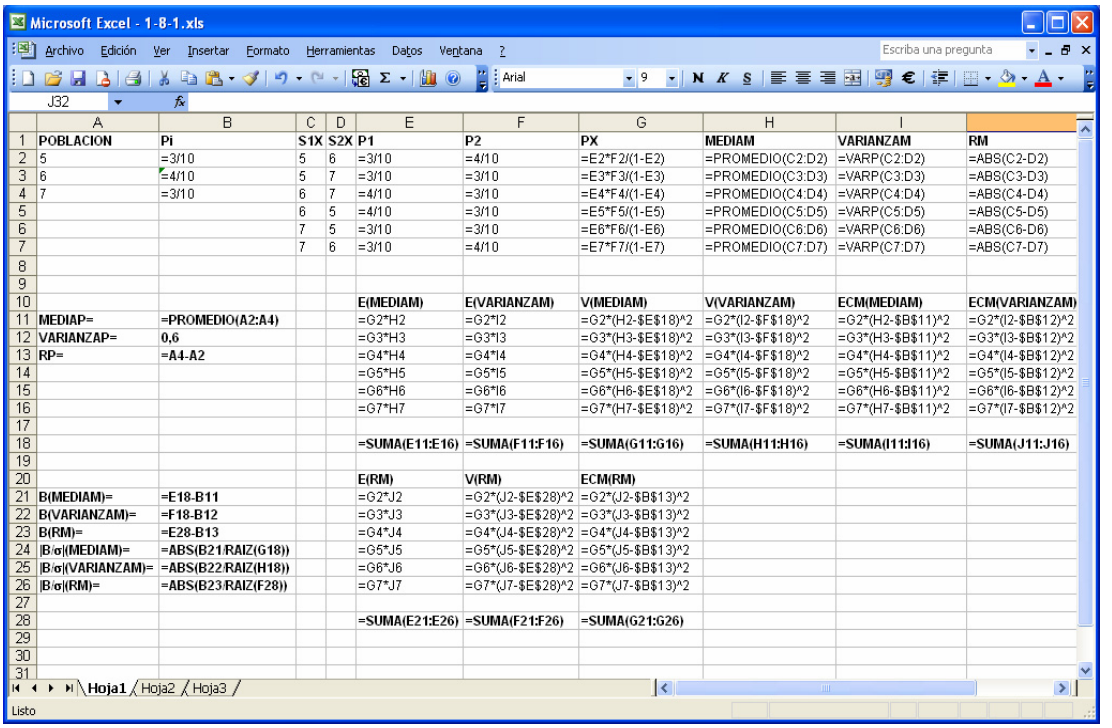
$$ECM(MEDIAM) = V(MEDIAM) = 0{,}185$$

$$ECM(VARIANZAM) = \sum_{i=1}^{6} \left(VARIANZAM_i - 0{,}6\right)^2 P_i = 0{,}132$$

$$ECM(MEDIAM) = V(MEDIAM) = 0{,}742$$

Como el estimador *MEDIAM* es insesgado, su varianza coincide con su error cuadrático medio, luego su precisión se mide a través de la varianza. De esta forma, el estimador *VARIANZAM* para estimar la varianza poblacional es más preciso que el estimador *MEDIAM* para estimar la media poblacional y que el estimador *RM* para estimar el recorrido poblacional por tener menor error cuadrático medio.

Los cálculos pueden implementarse mediante Excel tal y como se indica en las pantallas siguientes:



Microsoft Excel - 1-8-1.xls

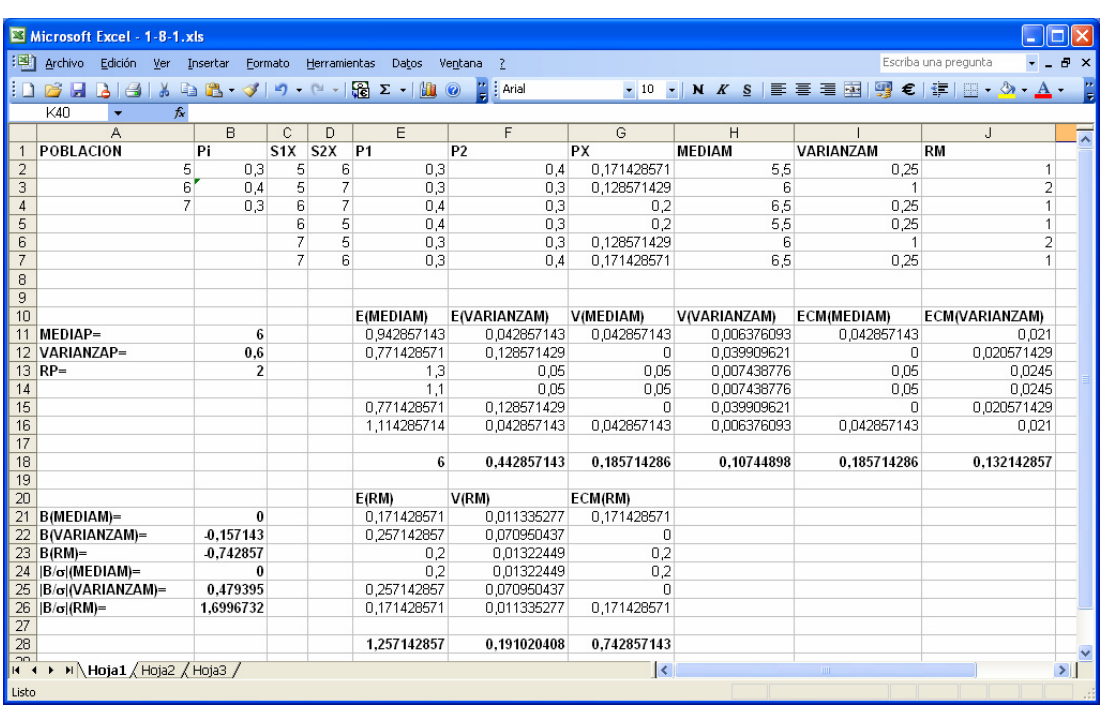
Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

J32

	A	B	C	D	E	F	G	H	I	
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM	RM
2	5	=3/10	5	6	=3/10	=4/10	=E2*F2/(1-E2)	=PROMEDIO(C2:D2)	=VARP(C2:D2)	=ABS(C2-D2)
3	6	=4/10	5	7	=3/10	=3/10	=E3*F3/(1-E3)	=PROMEDIO(C3:D3)	=VARP(C3:D3)	=ABS(C3-D3)
4	7	=3/10	6	7	=4/10	=3/10	=E4*F4/(1-E4)	=PROMEDIO(C4:D4)	=VARP(C4:D4)	=ABS(C4-D4)
5			6	5	=4/10	=3/10	=E5*F5/(1-E5)	=PROMEDIO(C5:D5)	=VARP(C5:D5)	=ABS(C5-D5)
6			7	5	=3/10	=3/10	=E6*F6/(1-E6)	=PROMEDIO(C6:D6)	=VARP(C6:D6)	=ABS(C6-D6)
7			7	6	=3/10	=4/10	=E7*F7/(1-E7)	=PROMEDIO(C7:D7)	=VARP(C7:D7)	=ABS(C7-D7)
8										
9										
10					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)	ECM(VARIANZAM)
11	MEDIAP=	=PROMEDIO(A2:A4)			=G2*H2	=G2*I2	=G2*(H2-$E$18)^2	=G2*(I2-$F$18)^2	=G2*(H2-$B$11)^2	=G2*(I2-$B$12)^2
12	VARIANZAP=	0,6			=G3*H3	=G3*I3	=G3*(H3-$E$18)^2	=G3*(I3-$F$18)^2	=G3*(H3-$B$11)^2	=G3*(I3-$B$12)^2
13	RP=	=A4-A2			=G4*H4	=G4*I4	=G4*(H4-$E$18)^2	=G4*(I4-$F$18)^2	=G4*(H4-$B$11)^2	=G4*(I4-$B$12)^2
14					=G5*H5	=G5*I5	=G5*(H5-$E$18)^2	=G5*(I5-$F$18)^2	=G5*(H5-$B$11)^2	=G5*(I5-$B$12)^2
15					=G6*H6	=G6*I6	=G6*(H6-$E$18)^2	=G6*(I6-$F$18)^2	=G6*(H6-$B$11)^2	=G6*(I6-$B$12)^2
16					=G7*H7	=G7*I7	=G7*(H7-$E$18)^2	=G7*(I7-$F$18)^2	=G7*(H7-$B$11)^2	=G7*(I7-$B$12)^2
17										
18					=SUMA(E11:E16)	=SUMA(F11:F16)	=SUMA(G11:G16)	=SUMA(H11:H16)	=SUMA(I11:I16)	=SUMA(J11:J16)
19										
20					E(RM)	V(RM)	ECM(RM)			
21	B(MEDIAM)=	=E18-B11			=G2*J2	=G2*(J2-$E$28)^2	=G2*(J2-$B$13)^2			
22	B(VARIANZAM)=	=F18-B12			=G3*J3	=G3*(J3-$E$28)^2	=G3*(J3-$B$13)^2			
23	B(RM)=	=E28-B13			=G4*J4	=G4*(J4-$E$28)^2	=G4*(J4-$B$13)^2			
24	|B/σ|(MEDIAM)=	=ABS(B21/RAIZ(G18))			=G5*J5	=G5*(J5-$E$28)^2	=G5*(J5-$B$13)^2			
25	|B/σ|(VARIANZAM)=	=ABS(B22/RAIZ(H18))			=G6*J6	=G6*(J6-$E$28)^2	=G6*(J6-$B$13)^2			
26	|B/σ|(RM)=	=ABS(B23/RAIZ(F28))			=G7*J7	=G7*(J7-$E$28)^2	=G7*(J7-$B$13)^2			
27										
28					=SUMA(E21:E26)	=SUMA(F21:F26)	=SUMA(G21:G26)			
29										
30										
31										

Hoja1 Hoja2 Hoja3

Listo



Microsoft Excel - 1-8-1.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

K40

	A	B	C	D	E	F	G	H	I	J
1	POBLACION	Pi	S1X	S2X	P1	P2	PX	MEDIAM	VARIANZAM	RM
2	5	0,3	5	6	0,3	0,4	0,171428571	5,5	0,25	1
3	6	0,4	5	7	0,3	0,3	0,128571429	6	1	2
4	7	0,3	6	7	0,4	0,3	0,2	6,5	0,25	1
5			6	5	0,4	0,3	0,2	5,5	0,25	1
6			7	5	0,3	0,3	0,128571429	6	1	2
7			7	6	0,3	0,4	0,171428571	6,5	0,25	1
8										
9										
10					E(MEDIAM)	E(VARIANZAM)	V(MEDIAM)	V(VARIANZAM)	ECM(MEDIAM)	ECM(VARIANZAM)
11	MEDIAP=	6			0,942857143	0,042857143	0,042857143	0,006376093	0,042857143	0,021
12	VARIANZAP=	0,6			0,771428571	0,128571429	0	0,039909621	0	0,020571429
13	RP=	2			1,3	0,05	0,05	0,007438776	0,05	0,0245
14					1,1	0,05	0,05	0,007438776	0,05	0,0245
15					0,771428571	0,128571429	0	0,039909621	0	0,020571429
16					1,114285714	0,042857143	0,042857143	0,006376093	0,042857143	0,021
17										
18					**6**	**0,442857143**	**0,185714286**	**0,10744898**	**0,185714286**	**0,132142857**
19										
20					E(RM)	V(RM)	ECM(RM)			
21	B(MEDIAM)=	0			0,171428571	0,011335277	0,171428571			
22	B(VARIANZAM)=	-0,157143			0,257142857	0,070950437	0			
23	B(RM)=	-0,742857			0,2	0,01322449	0,2			
24	|B/σ|(MEDIAM)=	0			0,2	0,01322449	0,2			
25	|B/σ|(VARIANZAM)=	0,479395			0,257142857	0,070950437	0			
26	|B/σ|(RM)=	1,6996732			0,171428571	0,011335277	0,171428571			
27										
28					**1,257142857**	**0,191020408**	**0,742857143**			

Hoja1 Hoja2 Hoja3

Listo

EJERCICIOS PROPUESTOS

1.1. Para la población $U=\{U_1, U_2, U_3\}$ consideramos el siguiente proceso de selección de muestras de tamaño 2. Se extrae una primera unidad con probabilidades iguales de selección, y si ésta resulta ser U_1, se extrae la segunda unidad entre las dos restantes también con probabilidades iguales; pero si la primera no es U_1, la segunda se extrae de las tres que componen la población asignando doble probabilidad a U_1 que a cada una de las otras dos. Hallar el espacio muestral y las probabilidades asociadas a las muestras para este procedimiento de muestreo. Si consideramos la variable X que toma los valores $X_i=\{1,1,0\}$ i = 1, 2, 3 en los tres elementos de la población y definimos el estimador para el total poblacional $\hat{X}=k(X_1+X_2)$, hallar su sesgo, su varianza y el valor de k para que sea insesgado.

1.2. Para medir la variable X = nivel de precipitación atmosférica en una determinada región disponemos de un marco de 4 zonas climáticas de la misma cuyos niveles de precipitación actual son de 6, 4, 3 y 8 decenas de litros por metro cuadrado, siendo sus probabilidades iniciales de selección en el muestreo 1/6, 1/3, 1/3 y 1/6, respectivamente. Se trata de estimar en decenas de litros por metro cuadrado el nivel actual medio de precipitación atmosférica en la región extrayendo muestras de la variable X con tamaño 2 sin reposición y sin tener en cuenta el orden de colocación de sus elementos. Para ello se consideran los estimadores alternativos MEDIA ARITMÉTICA, MEDIA GEOMÉTRICA, MEDIA CUADRÁTICA y MEDIA ARMÓNICA. Se pide lo siguiente:

1) Especificar el espacio muestral definido por este procedimiento de muestreo S(X), las probabilidades asociadas a las muestas P(S) y la distribución en el muestreo de los cuatro estimadores analizando su precisión. ¿Cuál de ellos es mejor? Razonar la respuesta y cuantificar las ganancias en precisión.

2) Hallar intervalos de confianza para la media según los cuatro estimadores basados en la muestra de mayor probabilidad para un nivel de confianza del 2 por mil (α=0,002). Como dato se sabe que $F^{-1}(0.999)= 3$, siendo F la función de distribución de la normal (0,1). Comentar los resultados.

1.3. Para la población $A=\{A_1, A_2, A_3, A_4, A_5\}$ consideramos el siguiente proceso de selección de muestras de tamaño 3. De una urna con tres bolas numeradas del 1 al 3 se extraen al azar y sin reposición dos bolas. A continuación, de otra urna con dos bolas numeradas con el 4 y el 5 se extrae una bola. Se pide:

1) Espacio muestral asociado a este experimento de muestreo y probabilidades de las muestras. Consideramos el estimador por analogía $\hat{\theta}$ = suma de los subíndices de unidades de las muestras para estimar la característica poblacional θ = suma de los subíndices de las unidades de población. Calcular la precisión del estimador y hallar un intervalo de confianza al 95%.

2) Se considera el estimador por analogía $\hat{\bar{\theta}}$ = Media de los subíndices de unidades de las muestras para estimar la característica poblacional $\hat{\theta}$ = Media de los subíndices de las unidades de población. Calcular la precisión de este estimador y hallar un intervalo de confianza al 95%. ¿Qué estimación es mejor? Cuantificar la ganancia en precisión.

1.4. Consideramos una población de 3 unidades $\{u_1, u_2, u_3\}$ cuyas probabilidades iniciales de selección son iguales a 1/3. Se extraen muestras de tamaño 2 con reposición sin tener en cuenta el orden de colocación de sus elementos. Se pide:

1) Espacio muestral y probabilidad asociadas a las muestras para este tipo de muestreo.

2) Se estima por analogía el parámetro poblacional θ = nº de unidades distintas en la población mediante el estimador $\hat{\theta}$ = nº de unidades distintas en la muestra Hallar la distribución en el muestreo del estimador $\hat{\theta}$ de θ.

3) Analizar la precisión de $\hat{\theta}$ para los valores θ= 1, θ= 2, θ= 3 del parámetro poblacional θ.

4) Se estima el parámetro poblacional $\overline{\theta}$ = Nº medio de unidades distintas en la población mediante el estimador por analogía $\hat{\overline{\theta}}$ = Nº medio de unidades distintas en la muestra. Hallar la distribución en el muestreo de $\hat{\overline{\theta}}$ y analizar su precisión para los valores $\overline{\theta}$ = 1 y $\overline{\theta}$ = 2 del parámetro poblacional $\overline{\theta}$.

5) ¿Cuál de las dos estimaciones anteriores es mejor? Hallar intervalos de confianza para ambos estimadores $\hat{\theta}$ y $\hat{\overline{\theta}}$ al 95% y comparar sus precisiones.

1.5. Para la población $A = \{A_1, A_2, \cdots, A_{12}\}$ consideramos el siguiente proceso de selección de muestras de tamaño 3. Se selecciona un entero al azar en el conjunto $\{1,2,3,4\}$ y siendo δ este número se forma la muestra $\{A_\delta, A_{\delta+4}, \cdots, A_{\delta+8}\}$. Considerando la variable $X_i=X(A_i)=i$ se pide la distribución, esperanza y varianza de los estimadores $T_1=\text{Máx}(X_i)$ y $T_2=2(\sum X_i)/n - 1$. ¿Cuál de los dos estimadores es más preciso? Realizar estimaciones por intervalos al 95% basadas en las muestras de mayor valor de los estimadores y comentar los resultados.

1.6. En una población con $N = 3$ unidades U_i (i = 1, 2, 3), la variable T_i medida sobre cada unidad toma los valores (1, 3, 5). Se considera un proceso de muestreo sin reposición con probabilidades iniciales de selección P_i = (1/5, 2/5, 2/5) y tamaño muestral $n = 2$ sin tener en cuenta el orden de colocación de las unidades en las muestras. Se pide:

1) Distribuciones en el muestreo de los estimadores $X=Ti+Tj$, $Y=Min(Ti,Tj)$, $Z=(Ti+Tj)/2$.
Si con X estimamos el total poblacional, con Y el menor valor de la población y con Z la media poblacional, ¿cuál de los tres estimadores es mejor? Razonar la respuesta y cuantificar las ganancias en precisión.

2) Hallar intervalos de confianza para los estimadores X, Y y Z basados en la muestra de mayor probabilidad para un nivel de confianza del 2 por mil ($F^{-1}(0.999)=3$ con $F \rightarrow N(0,1)$). Comentar los resultados.

1.7. Para la población $U = \{U_1, U_2, U_3\}$ se mide la variable X sobre sus unidades y se obtiene $X=\{3, 2, 4\}$. Se extrae una muestra de tamaño 2 mediante el siguiente proceso de selección. Se extraen dos bolas de una urna que tiene ocho (cuatro marcadas con un 1, tres con un 2 y una con un 3) y si sus números son (i,j) se extraen para la muestra las unidades (X_i, X_j). Hallar el espacio muestral, las probabilidades asociadas a las muestras y la distribución en el muestreo, esperanza y varianza del estimador por analogía media muestral. Resolver el problema para muestreo con y sin reposición.

CAPÍTULO

2

MÉTODOS GENERALES DE SELECCIÓN DE MUESTRAS. ESTIMACIÓN Y ERRORES

OBJETIVOS

1. Distinguir entre muestreo de unidades elementales y muestreo de unidades compuestas.
2. Distinguir claramente los conceptos de muestreo con probabilidades iguales y muestreo con probabilidades desiguales.
3. Distinguir entre muestreo con reposición y muestreo sin reposición.
4. Comprender cómo se forman los estimadores en el proceso de estimación puntual.
5. Comprender el concepto de factor de elevación.
6. Obtener el estimador lineal insesgado general para el caso de selección con reposición y probabilidades desiguales: Estimador de Hansen y Hurwitz.
7. Obtener la varianza y su estimación para el estimador de Hansen y Hurwitz.
8. Analizar los métodos especiales de selección con reposición y probabilidades desiguales: Método del tamaño acumulativo y método de Lahiri.
9. Obtener el estimador lineal insesgado general para el caso de selección sin reposición y probabilidades desiguales: Estimador de Horvitz y Thompson.
10. Obtener la varianza y la estimación de la varianza para el estimador de Horvitz y Thompson.
11. Obtener el estimador alternativo de Yates y Grundy para la varianza.
12. Analizar los métodos especiales de selección con reposición y probabilidades desiguales: Modelos de Ikeda, Mitzumo, Brewer, Durbin, Sampford y Murthy.
13. Analizar el muestreo con probabilidades gradualmente variables: Estimador de Sánchez Crespo y Gabeiras, error y estimación del error.
14. Obtener muestras aleatorias, especialmente mediante el método de Montecarlo.

ÍNDICE

1. Selección con y sin reposición. Probabilidades iguales y desiguales.

2. Estimación puntual y formación general de estimadores.

3. Muestreo con reposición y probabilidades desiguales. Estimador de Hansen Hurwitz.

4. Muestreo con reposición y probabilidades proporcionales a los tamaños. Métodos especiales de selección.

5. Muestreo sin reposición y probabilidades desiguales. Estimador de Horvitz y Thompson.

6. Muestreo sin reposición y probabilidades proporcionales a los tamaños. Métodos especiales de selección.

7. Método de Montecarlo

8. Problemas resueltos

9. Ejercicios propuestos

SELECCIÓN CON Y SIN REPOSICIÓN. PROBABILIDADES IGUALES Y DESIGUALES

Las formas básicas de selección de la muestra se clasifican atendiendo a los siguientes criterios:

1. *Atendiendo a las probabilidades de selección*

   1.1. *Con probabilidades iguales*: Todas las unidades de la población tienen la misma probabilidad de ser seleccionadas en cada extracción.

   1.2. *Con probabilidades desiguales*: Al menos dos unidades tienen distintas probabilidades de selección en cierta extracción.

2. *Atendiendo a la mecánica de selección*

   2.1. *Muestreo con reposición*: Cada unidad que es extraída para formar parte de la muestra en una extracción se repone a la población antes de realizar la siguiente extracción; es decir, la estructura poblacional permanece invariante.

   2.2. *Muestreo sin reposición*: Cada unidad que es extraída para formar parte de la muestra en una extracción no se repone a la población antes de realizar la siguiente extracción, por lo que una unidad podrá aparecer en la muestra a lo sumo una vez y la estructura poblacional va cambiando de una extracción a otra.

Combinando estos cuatro tipos de muestreo resulta:

- Muestreo con reposición y probabilidades iguales
- Muestreo sin reposición y probabilidades iguales
- Muestreo con reposición y probabilidades desiguales
- Muestreo sin reposición y probabilidades desiguales

ESTIMACIÓN PUNTUAL Y FORMACIÓN GENERAL DE ESTIMADORES

Supongamos que tenemos definida una característica X en la población $U = \{U_1, U_2, ..., U_N\}$ que toma el valor numérico X_i sobre la unidad U_i $i = 1, 2, ..., N$, dando lugar al conjunto de valores $\{X_1, X_2, ..., X_N\}$. Consideramos ahora una cierta función θ de los N valores X_i, que suele denominarse parámetro poblacional. Seleccionamos una muestra $s = \{u_1, u_2, ..., u_n\}$ de U mediante un procedimiento de muestreo dado, y consideramos los valores $s(X) = \{X_1, X_2, ... , X_n\}$ que toma la característica X en estudio sobre los elementos de la muestra. A partir de estos valores estimamos puntualmente el parámetro poblacional θ mediante la expresión $\hat{\theta} = \hat{\theta}(s(X)) = \hat{\theta}(X_1, ..., X_n)$, basada en los valores X_i $i = 1, 2, ..., n$, que toma la característica X sobre las unidades de la muestra s.

$$
\begin{array}{l}
U = \{U_1 \cdots U_N\} \xrightarrow{X} \left(X_1 \cdots X_N\right) \\
\downarrow \\
s = \{u_1 \cdots u_n\} \xrightarrow{X} s(X) = \left(X_1 \cdots X_n\right)
\end{array}
$$

La función $\hat{\theta}$ que asocia a cada muestra *s* el valor numérico $\hat{\theta}(s(X)) = \hat{\theta}(X_1, ..., X_n)$, se denomina *estimador* del parámetro poblacional θ. A los valores $\hat{\theta}(s(X))$ para cada *s* del espacio muestral se los denomina *estimaciones puntuales*. Por lo tanto podemos formalizar el concepto de estimador $\hat{\theta}$ para el parámetro poblacional θ definiéndolo mediante la aplicación medible:

$$\hat{\theta}: S(X) \subset R^n \rightarrow R$$
$$(X_1 \cdots X_n) \rightarrow \hat{\theta}(X_1 \cdots X_n) = t$$

Ya tenemos definido el estimador como un estadístico función de los valores que toma la característica X sobre los elementos del espacio muestral (muestras). Como ejemplos tenemos los estimadores total muestral y media muestral que estiman el total y la media poblacionales:

$$\hat{\theta}_1 : S(X) \subset R^n \rightarrow R$$
$$(X_1 \cdots X_n) \rightarrow \hat{\theta}_1(X_1 \cdots X_n) = X_1 + \cdots + X_n = \hat{X}$$

$$\hat{\theta}_2 : S(X) \subset R^n \rightarrow R$$
$$(X_1 \cdots X_n) \rightarrow \hat{\theta}_2(X_1 \cdots X_n) = \frac{X_1 + \cdots + X_n}{n} = \hat{\overline{X}}$$

Entre los parámetros poblacionales θ (función de los *N* valores poblacionales X_i) más comunes a estimar, tenemos el total poblacional y la media poblacional para la característica *X,* definidos de la forma siguiente:

- *Total poblacional:* $X = \theta(X_1, ..., X_N) = \sum_{i=1}^{N} X_i$
- *Media poblacional:* $\overline{X} = \theta(X_1, ..., X_N) = \frac{X}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i = \sum_{i=1}^{N} \frac{X_i}{N}$

Hasta ahora hemos supuesto que la característica *X* definida sobre los elementos de la población es cuantitativa, es decir, cuantificable numéricamente. Sin embargo, también se pueden definir características cualitativas sobre los elementos de la población, como, por ejemplo, su pertenencia o no a una determinada clase *A*. S_i para cada unidad u_i $i = 1, 2, ..., N$ de la población definimos la característica A_i, que toma valor 1 si la unidad u_i pertenece a la clase *A,* y que toma valor 0 si la unidad u_i no pertenece a la clase *A*, podemos definir el total de elementos de la población que pertenecen a la clase *A* (total de clase) y la proporción de elementos de la población que pertenecen a la clase *A* (proporción de clase) de la forma siguiente:

- *Total de clase:* $A = \theta(A_1, ..., A_N) = \sum_{i=1}^{N} A_i$
- *Proporción de clase:* $P = \theta(A_1, ..., A_N) = \frac{A}{N} = \frac{1}{N}\sum_{i=1}^{N} A_i = \sum_{i=1}^{N} \frac{A_i}{N}$

Analizados ya los cuatro parámetros poblacionales más típicos a estimar, vemos que, en general, un parámetro poblacional θ puede expresarse como una suma de elementos $Y_i = f(X_i)$ función de los valores que la característica cuantitativa *X* o cualitativa *A* considerada toma sobre los elementos de la población. De esta forma, podemos escribir:

$$\theta = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} f(X_i)$$

en cuyo caso tenemos:
$$\begin{cases} Y_i = f(X_i) = X_i \ \textit{para el total poblacional } X \\ Y_i = f(X_i) = \dfrac{X_i}{N} \ \textit{para la media poblacional } \overline{X} \\ Yi = f(A_i) = A_i \ \textit{para el total de clase } A \\ Yi = f(A_i) = \dfrac{A_i}{N} \ \textit{para la proporción de clase } P \end{cases}$$

Ahora surge el problema de analizar la forma de los estimadores puntuales óptimos $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ para estos parámetros poblacionales típicos. Resulta que las mejores propiedades suelen presentarlas los estimadores lineales insesgados de la forma $\hat{\theta} = \sum_{i=1}^{n} w_i Y_i$. Los valores ω_i se denominan ***pesos*** o ***factores de elevación***, ya que so los números por los que hay que multiplicar los valores muestrales para obtener los valores poblacionales.

Concretamente, *para muestreo sin reposición, el estimador óptimo es el de Horvitz y Thompson* $\hat{\theta}_{HT} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}$, donde π_i es la probabilidad que tiene la unidad u_i de la población de pertenecer a la muestra. Se observa que los pesos o factores de elevación son en este caso $w_i = \frac{1}{\pi_i}$.

Para muestreo con reposición el estimador óptimo es el de Hansen y Hurwitz $\hat{\theta}_{HH} = \sum_{i=1}^{n} \frac{Y_i}{nP_i}$, donde P_i es la probabilidad de seleccionar la unidad u_i de la población para la muestra (probabilidad unitaria de selección de la unidad u_i). Se observa que los pesos o factores de elevación son, en este caso, $w_i = \frac{1}{nP_i}$.

Existen justificaciones para considerar que el parámetro poblacional $\theta = \sum_{i=1}^{N} Y_i$ puede estimarse convenientemente mediante el estimador $\hat{\theta} = \sum_{i=1}^{n} w_i Y_i$, entre las que podemos citar:

- Todas las mediciones de la variable en estudio sobre las unidades de la muestra intervienen en la formación del estimador.
- La importancia de la aportación al estimador de la unidad muestral u_i puede controlarse mediante el coeficiente de ponderación w_i o factor de elevación.
- Cuando $w_i = 1$, todas las unidades muestrales intervienen de igual forma en la formación del estimador.
- Los coeficientes w_i pueden depender, entre otros factores, del tamaño de las unidades muestrales (cuando son compuestas), del orden de colocación de las mismas en la muestra, y sobre todo de la probabilidad que tiene la unidad u_i de pertenecer a la muestra según el método de muestreo considerado.
- Las funciones lineales son las más sencillas de manejar matemáticamente.

MUESTREO CON REPOSICIÓN Y PROBABILIDADES DESIGUALES: ESTIMADOR DE HANSEN HURWITZ

Consideremos una población de tamaño *N*, con unidades $\{u_1, u_2, \cdots, u_N\}$. Seleccionamos con reposición una muestra $(\tilde{\mathbf{x}})$ de tamaño *n*. Ya sabemos que en este esquema de selección cada unidad u_i de la población puede pertenecer a la muestra $(\tilde{\mathbf{x}})$ de tamaño *n* desde 0 a *n* veces ya que al seleccionar una unidad para la muestra, ésta se devuelve a la población antes de realizar la siguiente extracción.

La ***probabilidad de una muestra*** cualquiera de tamaño *n* seguirá el modelo multinomial (conjunta de *n* binomiales e_i), ya que al haber reposición puede seleccionarse para la muestra cada unidad u_i de la población t_i veces con $i = 1,2,...,N$ y $\sum_{i=1}^{N} t_i = n$, con lo que:

$$P(\tilde{x}) = P(\underbrace{u_1,\cdots u_1}_{t_1 veces},\underbrace{u_2,\cdots,u_2}_{t_2 veces},\cdots,\underbrace{u_N,\cdots,u_N}_{t_N veces}) = P(e_1 = t_1, e_2 = t_2,\cdots,e_N = t_N)$$

$$= \frac{n!}{t_1!t_2!\cdots t_N!} P_1^{t_1} P_2^{t_2} \cdots P_N^{t_N}\, n! = (t_1 + t_2 + \cdots + t_N)! \sum_{i=1}^{N} t_i = n$$

El estimador lineal insesgado óptimo en el muestreo con reposición y probabilidades desiguales para el parámetro poblacional $\theta = \sum_{i=1}^{N} Y_i$ es el ***estimador de Hansen y Hurwitz***:

$$\hat{\theta}_{HH} = \sum_{i=1}^{n} \omega_i Y_i = \sum_{i=1}^{n} \frac{1}{nP_i} Y_i = \sum_{i=1}^{n} \frac{Y_i}{nP_i}$$

Al particularizar el estimador de Hansen y Hurwitz para los distintos parámetros poblacionales, tenemos los siguientes estimadores:

Total → $\theta = X = \sum_{i=1}^{N} X_i \Rightarrow Y_i = X_i \Rightarrow \quad \hat{X}_{HH} = \sum_{i=1}^{n} \frac{X_i}{nP_i}$

Media → $\theta = \overline{X} = \sum_{i=1}^{N} \frac{X_i}{N} \Rightarrow Y_i = \frac{X_i}{N} \Rightarrow \quad \hat{\overline{X}}_{HH} = \sum_{i=1}^{n} \frac{\frac{X_i}{N}}{nP_i} = \frac{1}{N} \sum_{i=1}^{n} \frac{X_i}{nP_i}$

Total de clase → $\theta = A = \sum_{i=1}^{N} A_i \Rightarrow Y_i = A_i \Rightarrow \quad \hat{A}_{HH} = \sum_{i=1}^{n} \frac{A_i}{nP_i}$

Proporción→ $\theta = P = \sum_{i=1}^{N} \frac{A_i}{N} \Rightarrow Y_i = \frac{A_i}{N} \Rightarrow \quad \hat{P}_{HH} = \sum_{i=1}^{n} \frac{\frac{A_i}{N}}{nP_i} = \frac{1}{N} \sum_{i=1}^{n} \frac{A_i}{nP_i}$

Varianza del estimador de Hansen y Hurwitz

$$V(\hat{\theta}_{HH}) = \frac{1}{n} \sum_{i=1}^{N} \left(\frac{Y_i}{P_i} - \theta \right)^2 P_i = \frac{1}{n} \left[\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - \theta^2 \right] = \frac{1}{n} \sum_{i=1}^{N} \sum_{j>i}^{N} \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j} \right)^2 P_i P_j$$

Estimación de la varianza del estimador de Hansen y Hurwitz

$$\hat{V}(\hat{\theta}_{HH}) = \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{Y_i}{P_i}\right)^2 - n\hat{\theta}_{HH}^2\right] = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{Y_i}{P_i} - \hat{\theta}_{HH}\right)^2$$

SELECCIÓN CON REPOSICIÓN Y PROBABILIDADES PROPORCIONALES AL TAMAÑO: MÉTODOS ESPECIALES DE SELECCIÓN

Método de selección del tamaño acumulativo o modelo polinomial

Sea M_i un entero positivo asociado a la unidad u_i que denominamos tamaño de u_i para $i = 1, 2, ..., N$ (M_i puede ser el número de unidades elementales de la unidad compuesta u_i o una ponderación o medida de la importancia que concedemos a la selección de la unidad u_i para la muestra).

A continuación se expone un método práctico que permite seleccionar muestras con reposición de modo que en cada extracción la unidad u_i tiene probabilidad P_i proporcional a su tamaño M_i.

Sea $M = \sum_{i=1}^{N} M_i$. Consideramos el intervalo de números enteros [1, M] y lo dividimos en N subintervalos I_i cada uno de ellos con M_i unidades, tal y como se indica en el cuadro siguiente:

Subintervalos	*Unidades*	*Tamaños*
$I_1 = [1, M_1]$	u_1	M_1
$I_2 = [M_1 + 1, M_1 + M_2]$	u_2	M_2
$I_3 = [M_1 + M_2 + 1, M_1 + M_2 + M_3]$	u_3	M_3
$\vdots$	$\vdots$	$\vdots$
$I_N = [\left(\sum_{i=1}^{N-1} M_i\right) + 1, \underbrace{\sum_{i=1}^{N} M_i}_{M}]$	u_N	M_N

Ahora elegimos un entero $\delta \in [1,M]$ aleatoriamente y con probabilidades iguales y seleccionamos como primera unidad de la muestra la unidad u_i tal que $\delta \in I_i$. Repetimos este proceso n veces hasta obtener una muestra de tamaño n, de modo que para cualquiera de las n extracciones se cumple:

$$P(u_i) = P(\delta \in I_i) = \frac{M_i}{M} = P_i$$

El procedimiento de selección es con reposición, pues el entero $\delta \in [1,M]$ elegido aleatoriamente puede caer varias veces dentro del mismo intervalo I_i, con lo que la unidad u_i estará varias veces en la muestra. También hemos visto que el procedimiento de selección se realiza en cada extracción con probabilidades proporcionales a los tamaños, ya que $P_i=M_i/M$.

Este método también permite obtener muestras sin reposición. Basta no tener en cuenta la obtención de unidades repetidas y seguir seleccionando hasta acompletar el tamaño de muestra requerido. Por lo tanto, es un método general de selección de muestras.

Este método también permite ***extraer una muestrea con probabilidades desiguales no necesariamente proporcionales a sus tamaños***. Basta formar un rango acumulativo del los P_i y extraer una muestra de números aleatorios uniformes en (0,1). Es decir, basta montar un cuadro como el anterior donde los intervalos acumulativos Ii se formarían ahora con los $P_i = M_i/M$, en vez de con los M_i. Y en vez de obtener números aleatorios entre 1 y M, se obtendrían entre 0 y 1.

Método de selección de Lahiri

Una variante que abrevia el método del tamaño acumulativo la constituye el método de Lahiri, que permite también seleccionar muestras con reposición y probabilidades proporcionales a los tamaños.

Sea M_0 un número entero mayor o igual que todos los M_i, por ejemplo, $M_0 = \underbrace{Max}_{i=1,2,...,N}(M_i)$. Elegimos un par de números aleatorios (i, j) tales que $1 \le i \le N$ y $1 \le j \le M_0$.

Si $j \le M_i$, la unidad seleccionada para la muestra es la u_i. Si $j > M_i$ se repite la selección del par de números aleatorios (i, j) tales que $1 \le i \le N$ y $1 \le j \le M_0$ tantas veces como sea necesario hasta que $j \le M_i$.

Este método también permite obtener muestras sin reposición. Basta no tener en cuenta la obtención de unidades repetidas y seguir seleccionando hasta completar el tamaño de muestra requerido. Por lo tanto, es un método general de selección de muestras.

MUESTREO SIN REPOSICIÓN Y PROBABILIDADES DESIGUALES: ESTIMADOR DE HORVITZ THOMPSON

Decimos que un procedimiento aleatorio de muestreo es sin reposición cuando todas las muestras que tienen algún elemento repetido son imposibles. Las unidades seleccionadas no se reponen a la población para seleccionar la siguiente unidad de la muestra, con lo que las muestras resultantes tienen todos sus elementos distintos.

Decimos que un procedimiento aleatorio de muestreo es con probabilidades iguales cuando todas las unidades de la población u_i tienen la misma probabilidad de ser elegidas para la muestra en una determinada extracción. En caso de que no sea la misma estaremos ante muestreo con probabilidades desiguales. Tanto el muestreo con reposición como el muestreo sin reposición pueden ser con probabilidades iguales o desiguales.

En el caso de muestreo sin reposición y probabilidades desiguales, el estimador lineal insesgado para el parámetro poblacional $\theta = \sum_{i=1}^{N} Y_i$ es el ***estimador de Horvitz y Thompson***:

$$\hat{\theta}_{HT} = \sum_{i=1}^{n} w_i Y_i = \sum_{i=1}^{n} \frac{1}{\pi_i} Y_i = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}$$

donde π_i es la probabilidad de que la unidad i de la población pertenezca a la muestra.

Al particularizar el estimador de Horvitz y Thompson para los distintos parámetros poblacionales, tenemos los siguientes estimadores:

Total → $$\theta = X = \sum_{i=1}^{N} X_i \Rightarrow Y_i = X_i \Rightarrow \hat{X}_{HT} = \sum_{i=1}^{n} \frac{X_i}{\pi_i}$$

Media → $$\theta = \bar{X} = \sum_{i=1}^{N} \frac{X_i}{N} \Rightarrow Y_i = \frac{X_i}{N} \Rightarrow \hat{\bar{X}}_{HT} = \frac{1}{N} \sum_{i=1}^{n} \frac{X_i}{\pi_i}$$

Total de clase → $$\theta = A = \sum_{i=1}^{N} A_i \Rightarrow Y_i = A_i \Rightarrow \hat{A}_{HT} = \sum_{i=1}^{n} \frac{A_i}{\pi_i}$$

Proporción→ $$\theta = P = \sum_{i=1}^{N} \frac{A_i}{N} \Rightarrow Y_i = \frac{A_i}{N} \Rightarrow \hat{P}_{HT} = \frac{1}{N} \sum_{i=1}^{n} \frac{A_i}{\pi_i}$$

Varianza del estimador de Horvitz y Thompson

$$V\left(\hat{\theta}_{HT}\right) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}\left(1-\pi_i\right) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\left(\pi_{ij} - \pi_i\pi_j\right)$$

donde π_i es la probabilidad de que la unidad *i* de la población pertenezca a la muestra y π_{ij} es la probabilidad de que el par de unidades de la población (*i,j*) pertenezcan a la muestra.

Estimación de la varianza del estimador de Horvitz y Thompson

$$\hat{V}\left(\hat{\theta}_{HT}\right) = \sum_{i=1}^{n} \frac{Y_i^2}{\pi_i^2}\left(1-\pi_i\right) + 2\sum_{i=1}^{n}\sum_{j>i}^{n} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\frac{\left(\pi_{ij} - \pi_i\right)}{\pi_{ij}}$$

Estimador de la varianza de Yates y Grundy para el estimador de Horvitz y Thompson

$$\hat{V}\left(\hat{\theta}_{HT}\right) = \sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \frac{\left(\pi_i\pi_j - \pi_{ij}\right)}{\pi_{ij}}$$

Todas las fórmulas para estimadores y errores vistas hasta ahora para el caso general sin reposición con probabilidades desiguales son válidas para el ***caso particular de probabilidades iguales sin reposición*** haciendo las siguientes sustituciones:

$$\pi_i = \frac{n}{N},\ \pi_{ij} = \frac{n(n-1)}{N(N-1)}$$

Se observa que cualquier método de selección sin reposición queda perfectamente definido al conocer π_i y π_{ij} ya que los estimadores y sus errores dependen sólo de estos valores.

SELECCIÓN SIN REPOSICIÓN Y PROBABILIDADES PROPORCIONALES AL TAMAÑO: MÉTODOS ESPECIALES DE SELECCIÓN

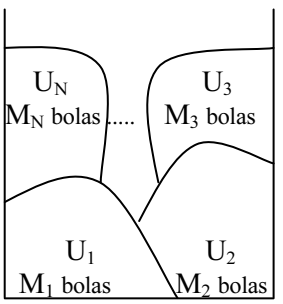
Estimador de Horvitz y Thompson con probabilidades proporcionales al tamaño

Sea M_i el entero positivo asociado a la unidad compuesta U_i $i = 1,\ldots,N$ que representa su tamaño (número de unidades elementales que contiene). En la práctica las unidades de muestreo suelen ser conglomerados, aunque a veces este modelo también suele utilizarse con unidades de muestreo simples, en cuyo caso los M_i son ponderaciones utilizadas para dar un mayor peso o importancia a determinadas unidades muestrales.

Mediante este modelo se selecciona sin reposición de la población para la muestra la unidad compuesta U_i de tamaño M_i. Como se trata del modelo clásico de selección sin reposición, se procede a retirar de la población las M_i unidades elementales que componen la unidad de muestreo compuesta U_i antes de proceder a la selección para la muestra de la siguiente unidad de muestreo compuesta. Cuando se realiza la siguiente selección ya faltan de la población M_i unidades elementales. Se supone que en la población hay N unidades de muestreo compuestas que contienen un total de M unidades elementales, es decir:

$$M = \sum_{i=1}^{N} M_i$$

Este modelo clásico de selección de la muestra sin reposición es equivalente a considerar un ***modelo de urna generalizado*** consistente en introducir en una urna M bolas que representan las unidades elementales de la población y que se clasifican en N grupos distinguibles, cada uno de los cuales tiene las M_i bolas correspondientes al tamaño de la unidad compuesta U_i, de tal forma que cada unidad compuesta de muestreo U_i queda representada en la urna por M_i bolas distinguibles. Si en una extracción se obtiene una bola que representa una unidad elemental del grupo de la unidad compuesta U_i, se procede a retirar de la urna las M_i bolas correspondientes a todas las unidades elementales de U_i antes de realizar la siguiente selección.



Según este modelo, la probabilidad de seleccionar la unidad U_i en una extracción (probabilidad unitaria de selección) es $P_i = M_i/M = p(u_i)$ $i = 1, 2,..., N$. Se cumple que:

$$P_i = \frac{M_i}{M} = \frac{M_i}{\sum_{i=1}^{N} M_i} \Rightarrow \sum_{i=1}^{N} P_i = \sum_{i=1}^{N} \frac{M_i}{M} = \frac{\sum_{i=1}^{N} M_i}{M} = \frac{M}{M} = 1$$

con lo que el modelo está bien definido. Los valores π_i y π_{ij} relativos, respectivamente, a la probabilidad de que una unidad de la población pertenezca a la muestra y de que un par de unidades de la población pertenezcan a la muestra para muestras de tamaño 2 son:

$$\pi_i = P_i\left(1 + \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{P_j}{1-P_j}\right) = P_i\left(\frac{1-2P_i}{1-P_i} + \sum_{i=1}^{N} \frac{P_i}{1-P_i}\right) \qquad \pi_{ij} = P_iP_j\left[\frac{1}{1-P_i} + \frac{1}{1-P_j}\right]$$

Al conocer π_i y π_{ij} este método de selección sin reposición queda perfectamente definido, ya que los estimadores y sus errores dependen sólo de estos valores.

El método es generalizable para muestras de tamaño n.

Estimador de Horvitz y Thompson con selección Brewer

Brewer propuso un método de selección para muestras de tamaño *n* tal que la primera unidad se extrae sin reposición con probabilidad proporcional al valor:

$$k_i = P_i \frac{(1-P_i)}{(1-2P_i)} \qquad P_i < ½$$

y el resto de las extracciones se realizan sin reposición y con probabilidades proporcionales a P_i.

Para muestras de tamaño 2 las probabilidades π_i y π_{ij} son las siguientes:

$$\pi_i = 2P_i \qquad \pi_{ij} = \frac{2P_iP_j}{1+\sum_{i=1}^{N}\frac{P_i}{1-2P_i}} * \left[\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right]$$

Para muestras de tamaño *n* se demuestra que $\pi_i = nP_i$.

Estimador de Horvitz y Thompson con selección Durbin

El método de Durbin consiste en un muestreo con probabilidades desiguales y sin reemplazamiento con el siguiente método de selección para una muestra de tamaño $n = 2$: la primera unidad es seleccionada con probabilidad dada P_i y la segunda unidad se selecciona con probabilidades proporcionales a k_j, siendo:

$$k_j = P_j\left[\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right]$$

Para muestras de tamaño 2 las probabilidades π_i y π_{ij} son las siguientes:

$$\pi_i = 2P_i \qquad \pi_{ij} = \frac{2P_iP_j}{1+\sum_{i=1}^{N}\frac{P_i}{1-2P_i}} * \left[\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right]$$

Para muestras de tamaño *n* se demuestra que $\pi_i = nP_i$.

Se observa que los valores de π_i y π_{ij} son idénticos a los obtenidos con el método de selección de Brewer. Con estos valores ya se pueden hallar estimadores y errores.

Estimador de Horvitz y Thompson con selecciones de Ikeda y Mitzuno

Ikeda propuso un método de selección en el que la primera unidad se obtiene sin reposición con probabilidad P_i proporcional a su tamaño M_i y las $n-1$ unidades restantes de la muestra se seleccionan sin reposición y con probabilidades iguales.

Los valores de π_i y π_{ij} para este método son:

$$\pi_i = P_i + (1-P_i) * \frac{n-1}{N-1} = \frac{N-n}{N-1} * P_i + \frac{n-1}{N-1}$$

$$\pi_{ij} = P_i * \frac{n-1}{N-1} + P_j\frac{n-1}{N-1} + (1-(P_i+P_j)) * \frac{n-1}{N-1} * \frac{n-2}{N-2} = \frac{n-1}{N-1} * \left[\frac{N-n}{N-2}(P_i+P_j)+\frac{n-2}{N-2}\right]$$

Este método de Ikeda es un caso particular del ***método más general de Mitzuno***, que consiste en comenzar efectuando m extracciones sin reposición y con probabilidades iguales; en la extracción $m + 1$ se asignan probabilidades:

$$P_i + \sum_{r=1}^{m} \frac{P_r}{N-m}$$

donde P_r corresponde a la unidad extraída en r-ésimo lugar ($1 \leq r \leq m$), y por último las $n-(m+1)$ unidades muestrales restantes se seleccionan sin reposición y probabilidades iguales. El método de Ikeda es un caso particular del método de Mitzuno para $m = 0$.

Estimador de Horvitz y Thompson con selección Sampford

En este método los elementos muestrales se eligen con reposición seleccionando el primer elemento con probabilidad P_i y los restantes $n - 1$ elementos con probabilidades proporcionales a $P_i/(1 - nP_i)$. Finalizada la extracción, la muestra se acepta si todos los elementos son diferentes, y en caso contrario se rechaza y se vuelve a empezar. Se tiene que:

$$\pi_i = nP_i \qquad \pi_{ij} \approx n(n-1)P_iP_j \left(\begin{array}{l} 1 + \left[\left(P_i + P_j\right) - \sum_k P_k^2 \right] + 2\left(P_i^2 + P_j^2\right) - 2\sum_k P_k^3 - (n-2)P_iP_j + \\ + (n-3)\left(P_i + P_j\right) - \sum_k P_k^3 - (n-3)\left(\sum_k P_k^2\right) \end{array} \right)$$

Muestreo con probabilidades gradualmente variables

Se considera un esquema de urna en el que la unidad U_i viene representada por M_i bolas. En este esquema de selección con ***probabilidades gradualmente variables***, al seleccionar la unidad U_i se retira una bola de entre las M_i que representan a U_i y ***no se vuelve a reponer a la urna para la siguiente extracción***. Se podrá extraer la unidad U_i las veces que corresponda mientras no se acaben las M_i bolas que la representan o mientras no se cubra el tamaño n de la muestra, por lo que la unidad U_i puede figurar en la muestra un máximo de veces igual a $Mín(M_i,n)$ i=1,..,N.

La ***probabilidad de una muestra*** de tamaño n seguirá el modelo hipergeométrico generalizado (conjunta de n hipergeométricas e_i). Si cada unidad U_i de la población puede elegirse para la muestra t_i veces con $i = 1, 2, ..., N$ y se cumple que $\sum_{i=1}^{N} t_i = n$, tenemos:

$$P(\tilde{x}) = P(\underbrace{U_1,\cdots U_1}_{t_1 veces},\underbrace{U_2,\cdots,U_2}_{t_2 veces},\cdots,\underbrace{U_N,\cdots,U_N}_{t_N veces}) = P(e_1 = t_1, e_2 = t_2,\cdots,e_N = t_N)$$

$$= \frac{\binom{M_1}{t_1}\binom{M_2}{t_2}\cdots\binom{M_N}{t_N}}{\binom{M_1 + M_2 + \cdots M_N}{t_1 + t_2 + \cdots t_N}} = \frac{\binom{M\cdot P_1}{t_1}\binom{M\cdot P_2}{t_2}\cdots\binom{M\cdot P_N}{t_N}}{\binom{M}{n}} \quad con \quad \sum_{i=1}^{N} t_i = n$$

Mediante selección con probabilidades gradualmente variables, el estimador lineal insesgado (de Sánchez Crespo y Gabeiras) para el parámetro poblacional $\theta = \sum_{i=1}^{N} Y_i$ será:

$$\boxed{\hat{\theta}_{SCG} = \sum_{i=1}^{n} \omega_i Y_i = \sum_{i=1}^{n} \frac{1}{nP_i} Y_i = \sum_{i=1}^{n} \frac{Y_i}{nP_i} = \hat{\theta}_{HH}}$$

que coincide con la expresión del estimador de Hansen y Hurwitz para muestreo con reposición y probabilidades desiguales. Se cumple que:

$$V\left(\hat{\theta}_{SCG}\right)=\frac{M-n}{M-1}V(\hat{\theta}_{HH})$$

$$\hat{V}(\hat{\theta}_{SCG})=\frac{M-n}{M}\frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{Y_i}{P_i}\right)^2-n\hat{\theta}_{SCG}^2\right]=\frac{M-n}{M}\hat{V}(\hat{\theta}_{HH})$$

Se observa que el estimador de Sánchez Crespo y Gabeiras tiene menor varianza y menor varianza estimada que el estimador de Hansen y Hurwitz, ya que:

$$V\left(\hat{\theta}_{SCG}\right)=\frac{M-n}{M-1}V(\hat{\theta}_{HH})\le V(\hat{\theta}_{HH})\quad\text{y}\quad\hat{V}\left(\hat{\theta}_{SCG}\right)=\frac{M-n}{M}\hat{V}(\hat{\theta}_{HH})\le\hat{V}(\hat{\theta}_{HH})$$

Gabeiras sugirió una ***generalización del método anterior*** consistente en retirar b bolas en lugar de una cuando la unidad *i*-ésima es seleccionada para formar parte de la muestra, supuesto un esquema de urnas en el que la unidad U_i está representada por M_i bolas $(i=1,...,N)$, siendo b el mayor valor que permita a todas las unidades estar representadas en la urna durante las n extracciones, es decir, $b=\dfrac{Min(M_i)}{n-1}$.

Sánchez Crespo comprobó que con esta restricción la varianza del estimador resultante es menor e incluso en ciertos casos inferior a la varianza de los estimadores obtenidos bajo un muestreo sin reposición y probabilidades desiguales. La varianza del estimador para el total con el esquema mixto (generalización del muestreo gradual) viene dada por la expresión:

$$V(\hat{X}_{SC})=\frac{M-bn}{M-b}V(\hat{X}_{HH})$$

Se denomina ***esquema mixto*** a este procedimiento de muestreo ya que puede considerarse con reposición, en el sentido de que cada unidad puede pertenecer a la muestra más de una vez, y sin reposición, pues no se reponen en la urna las b unidades retiradas en cada extracción.

Método de Murthy

Murthy mejoró un método anterior de Des Raj extrayendo unidades sucesivas para la muestra con probabilidades P_i, $P_j(1-P_i)$, $P_k(1-P_i-P_j)$ y así sucesivamente. Propuso el estimador del total:

$$\hat{X}_M=\frac{\sum_{i=1}^{n}P(S/i)X_i}{P(S)},\quad\hat{V}\left(\hat{X}_M\right)=\frac{1}{P(S)^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\left[P(S)P(S/i,j)-P(S/i)P(S/j)\right]P_iP_j\left(\frac{X_i}{P_i}-\frac{X_j}{P_j}\right)^2$$

P(S) = Probabilidad incondicional de obtener la muestra *S*.
P(S/i) = Probabilidad de obtener la muestra *S* condicionado a que se sacó la unidad *i* la primera
P(S/i,j) = Probabilidad de *S* condicionado a que se sacaron las unidades *i* y *j* las dos primeras.

Para n=2 se tiene que $P(S/i)=P_j/(1-P_i)$ y $P(S)=\pi_{ij}=P_iP_j(2-P_i-P_j)/(1-P_i)(1-P_j)$ y además:

$$\pi_i=P_i\left[1+\sum_{j\neq i}\frac{P_j}{1-P_j}\right]\quad\hat{X}_M=\frac{1}{2-P_i-P_j}\left[(1-P_j)\frac{X_i}{P_i}+(1-P_i)\frac{X_j}{P_j}\right],\quad\hat{V}(\hat{X}_M)=\frac{(1-P_i)(1-P_j)(1-P_i-P_j)}{(2-P_i-P_j)^2}\left(\frac{X_i}{P_i}-\frac{X_j}{P_j}\right)^2$$

MÉTODO DE MONTECARLO

Es un procedimiento general para seleccionar muestras aleatorias simples de cualquier población (finita o infinita, real o teórica) de la que se conoce su distribución de probabilidad.

Variable aleatoria discreta

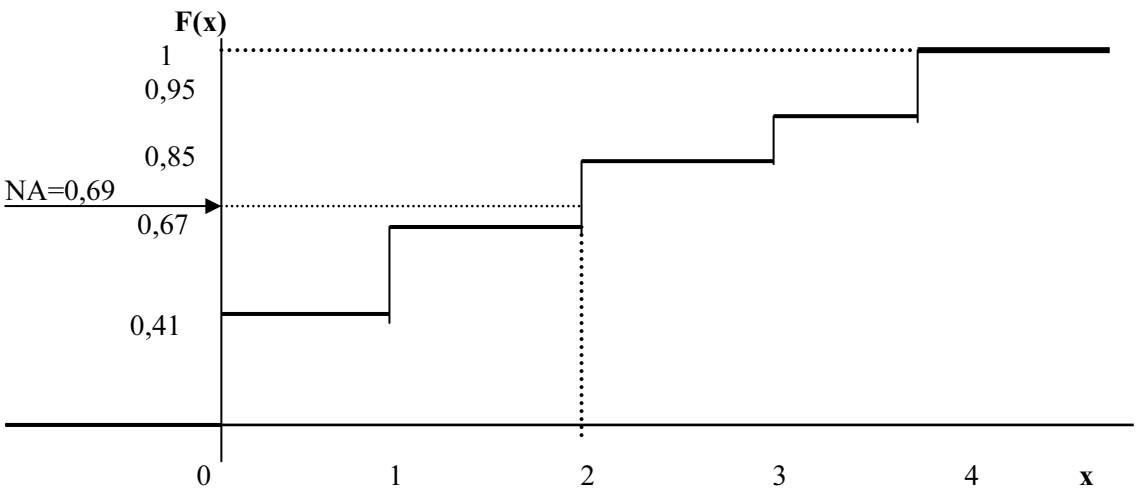
Consideremos la variable aleatoria discreta siguiente y veremos cómo se toma una muestra de ella.

x	$P(x)$	$F(x)$
0	0,41	0,41
1	0,26	0,67
2	0,18	0,85
3	0,10	0,95
4	0,05	1

Observamos los valores de la función de distribución y, basándonos en ellos, construimos la tabla:

Intervalos	x	$F(x)$
00-40	0	0,41
41-66	1	0,67
67-84	2	0,85
85-94	3	0,95
95-99	4	1

Para seleccionar la muestra aleatoria según la variable X, elegimos un número aleatorio entre 0 y 99 y observamos en qué intervalo cae, eligiendo para la muestra el valor x correspondiente a ese intervalo. También se puede tomar el número aleatorio y convertirlo en decimal NA (por ejemplo, si sale 69 utilizamos NA = 0,69) y tomar para la muestra el valor x más pequeño que verifica $F(x) > \text{NA}$

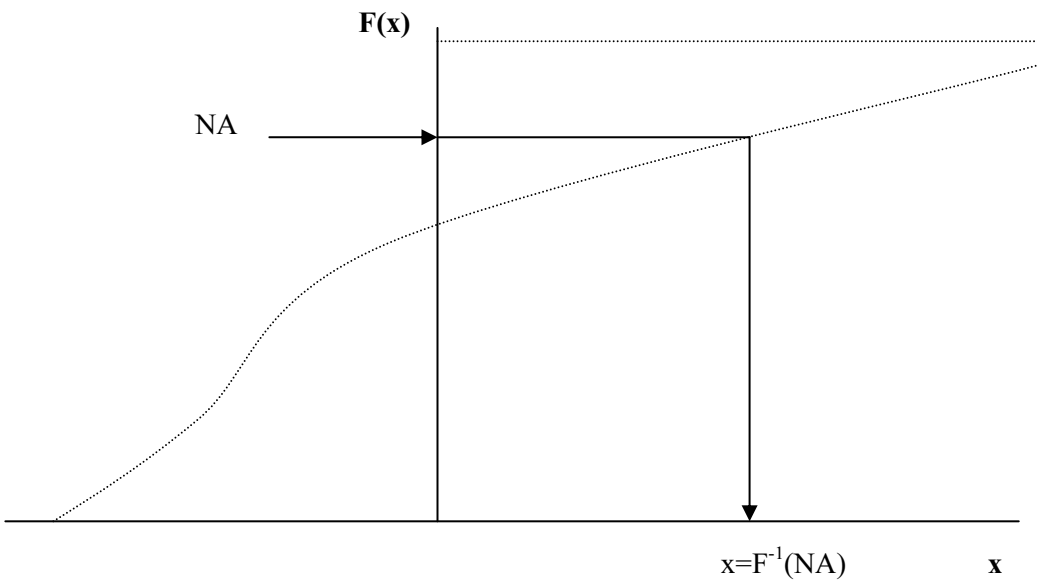


El valor x más pequeño que verifica $F(x) > 0{,}69$ es $x = 2$, luego el primer valor para la muestra es $x = 2$.

Variable aleatoria continua

El proceso a seguir sería el siguiente:

- Tomar un número aleatorio de tantas cifras como precisión se desee y convertirlo en decimal (por ejemplo, 23457 se convertiría en 0,23457), y sea NA dicho valor.
- Considerar el valor NA como un valor de $F(x)$ y tomar como valor observado en la muestra aquel valor de x tal que $NA=F(x) \Rightarrow x=F^{-1}(NA)$.
- Repetir el proceso con distintos números aleatorios hasta completar el tamaño de muestra deseado.



Dado el número aleatoria NA, se toma para la muestra el valor x tal que $x=F^{-1}(NA)$.

PROBLEMAS RESUELTOS

2.1. Un investigador desea muestrear tres hospitales de entre los seis que existen en una ciudad, con el propósito de estimar la proporción de pacientes que han estado (o estarán) en el hospital por más de dos días consecutivos. Puesto que los hospitales varían en tamaño, éstos serán muestreados con probabilidades proporcionales al número de sus pacientes. Con la información sobre los hospitales dada en la tabla adjunta se selecciona una muestra de tres hospitales con probabilidades proporcionales al tamaño (número de pacientes) con reposición utilizando el modelo del tamaño acumulativo (o modelo polinomial).

Hospital	Pacientes	Hospital	Pacientes	Hospital	Pacientes
1	328	2	109	3	432
4	220	5	280	6	190

Puesto que serán seleccionados tres hospitales, deben ser elegidos tres números aleatorios entre el 0001 y el 1559 = ∑Pacientes. Nuestros números elegidos son 1505, 1256 y 0827. ¿Qué hospitales serán elegidos para la muestra? Supóngase que los hospitales muestreados registraron los siguientes datos sobre el número de pacientes con permanencia de más de dos días:

Hospital	Nº de pacientes con más de dos días de permanencia
a	250
b	150
c	80

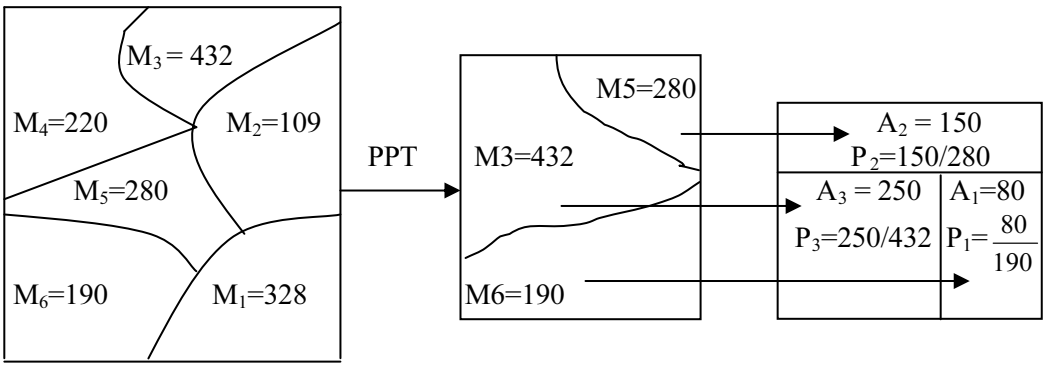
1) Estimar la proporción de pacientes con permanencia superior a dos días para los seis hospitales.

2) Establecer un límite para el error de estimación con una confianza del 95%.

Para seleccionar la muestra comenzamos construyendo la tabla relativa al método del tamaño acumulativo.

$I_1 = [1, M_1] = [1{,}328]$	*Unidades*	*Tamaños*
$I_2 = [M_1 + 1, M_1 + M_2] = [329{,}437]$	u_1	M_1
$I_3 = [438{,}869] \rightarrow \underline{827}$	u_2	M_2
$I_4 = [870{,}1089]$	u_3	M_3
$I_5 = [1090{,}1369] \rightarrow \underline{1256}$	⋮	⋮
$I_6 = [1370{,}1559] \rightarrow \underline{1505}$	u_N	M_N

Para seleccionar tres hospitales para la muestra se eligen tres números aleatorios entre 0001 y 1559 que resultan ser el 1505, el 1256 y el 0827. Localizados estos números en la columna de los intervalos acumulados, seleccionamos para la muestra los hospitales 3, 5 y 6. A continuación se presenta un esquema ilustrativo de la selección de las unidades muestrales.



N = 6, M=1559 n = 3

A continuación se realiza la estimación de la proporción de pacientes con permanencia superior a dos días utilizando el estimador de Hansen y Hurwitz (ya que el método de selección de la muestra es con reposición). Se tiene:

$$\hat{\bar{X}} = \frac{1}{M}\hat{X}_{HH} = \frac{1}{M}\sum_i^n \frac{X_i}{nP_i} = \frac{1}{M}\sum_i^n \frac{M_i\bar{X}_i}{n\frac{M_i}{M}} = \frac{1}{n}\sum_i^n \bar{X}_i \Rightarrow \hat{P} = \frac{1}{n}\sum_i^n \hat{P}_i = \frac{1}{3}\left(\frac{80}{190}+\frac{150}{280}+\frac{250}{432}\right) = 0{,}51$$

Por lo tanto, se estima que un 51% de los pacientes permanece más de dos días en el hospital. A continuación hallamos el error de esta estimación.

$$\hat{V}(\hat{\theta}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^n\left(\frac{Y_i}{P_i}-\hat{\theta}_{HH}\right)^2 \Rightarrow \hat{V}(\hat{\bar{X}}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^n\left(\frac{\frac{X_i}{M}}{\frac{M_i}{M}}-\hat{\bar{X}}_{HH}\right)^2 = \frac{1}{n(n-1)}\left(\sum_{i=1}^n \bar{X}_i \ -\hat{\bar{X}}_{HH}\right)^2$$

con lo que se tiene:

$$\hat{V}\left(\hat{P}\right) = \frac{\sum_i^n\left(\hat{P}_i-\hat{P}\right)^2}{n(n-1)} = \frac{1}{3\cdot 2}\left[\left(\frac{80}{190}-0{,}51\right)^2+\left(\frac{150}{280}-0{,}51\right)^2+\left(\frac{250}{432}-0{,}51\right)^2\right] = 0{,}0022$$

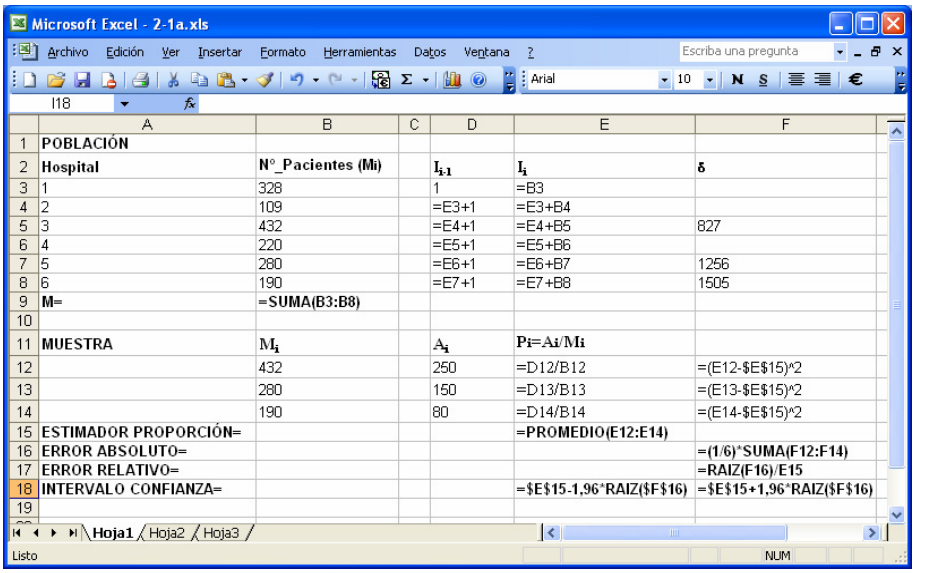
$$\hat{C}_v\left(\hat{P}\right) = \frac{\sqrt{\hat{V}\left(\hat{P}\right)}}{\hat{P}} = \frac{\sqrt{0{,}0022}}{0{,}51} = 0{,}0091 \rightarrow 1\%$$

Se observa que el error relativo de muestreo es del 1%. A continuación se realiza una estimación por intervalos al 95% de confianza.

$$\hat{P} \pm \lambda_\alpha\sqrt{\hat{V}(\hat{P})} = 0{,}51 \pm 1{,}96\sqrt{0{,}0022} = [0.4,\ 0.6] \rightarrow 95\%\ \textit{confianza}$$

Se observa que el intervalo de confianza es muy estrecho. Esto se debe a que la estimación realizada es bastante precisa (solamente un 5% de error).

Utilizando la hoja de cálculo Excel, se pueden automatizar los cálculos anteriores tal y como se indica en las figuras siguientes (en las figuras, P_i juega el papel de $\hat{P}_i$)



Microsoft Excel - 2-1a.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

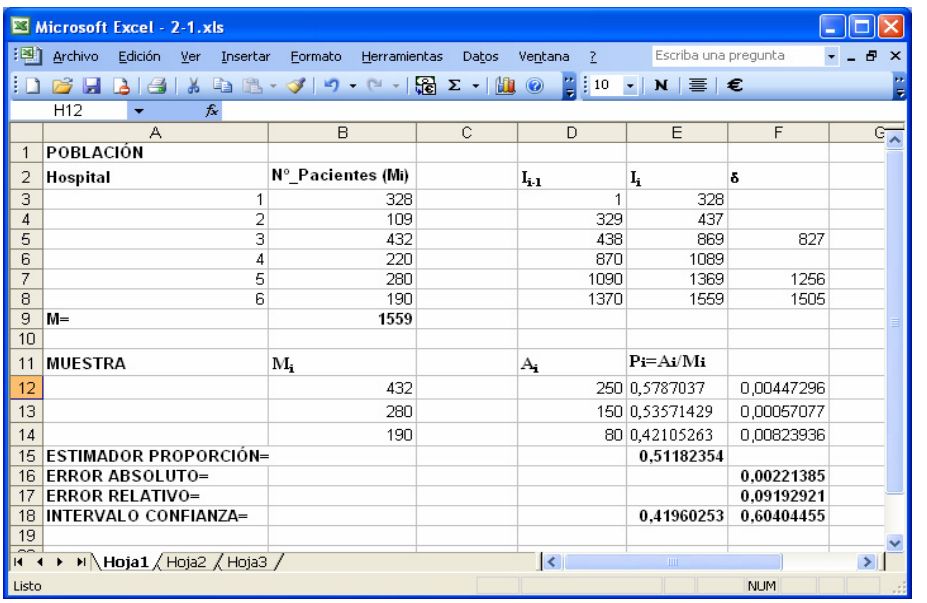
Arial 10

I18

	A	B	C	D	E	F
1	POBLACIÓN					
2	Hospital	N°_Pacientes (Mi)		I_{i-1}	I_i	δ
3	1	328		1	=B3	
4	2	109		=E3+1	=E3+B4	
5	3	432		=E4+1	=E4+B5	827
6	4	220		=E5+1	=E5+B6	
7	5	280		=E6+1	=E6+B7	1256
8	6	190		=E7+1	=E7+B8	1505
9	M=	=SUMA(B3:B8)				
10						
11	MUESTRA	M_i		A_i	Pi=Ai/Mi	
12		432		250	=D12/B12	=(E12-$E$15)^2
13		280		150	=D13/B13	=(E13-$E$15)^2
14		190		80	=D14/B14	=(E14-$E$15)^2
15	ESTIMADOR PROPORCIÓN=				=PROMEDIO(E12:E14)	
16	ERROR ABSOLUTO=					=(1/6)*SUMA(F12:F14)
17	ERROR RELATIVO=					=RAIZ(F16)/E15
18	INTERVALO CONFIANZA=				=$E$15-1,96*RAIZ($F$16)	=$E$15+1,96*RAIZ($F$16)
19						

Hoja1 Hoja2 Hoja3

Listo NUM



Microsoft Excel - 2-1.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

Escriba una pregunta

10

H12

	A	B	C	D	E	F	G
1	POBLACIÓN						
2	Hospital	N°_Pacientes (Mi)		I_{i-1}	I_i	δ	
3	1	328		1	328		
4	2	109		329	437		
5	3	432		438	869	827	
6	4	220		870	1089		
7	5	280		1090	1369	1256	
8	6	190		1370	1559	1505	
9	M=	1559					
10							
11	MUESTRA	M_i		A_i	Pi=Ai/Mi		
12		432		250	0,5787037	0,00447296	
13		280		150	0,53571429	0,00057077	
14		190		80	0,42105263	0,00823936	
15	ESTIMADOR PROPORCIÓN=				0,51182354		
16	ERROR ABSOLUTO=					0,00221385	
17	ERROR RELATIVO=					0,09192921	
18	INTERVALO CONFIANZA=				0,41960253	0,60404455	
19							

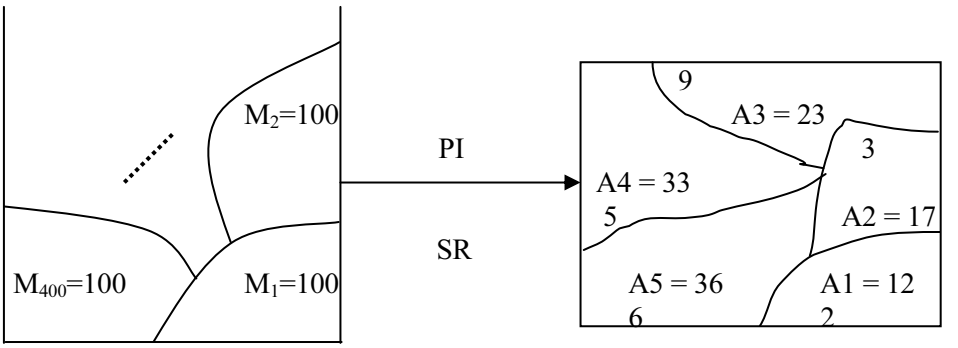
Hoja1 Hoja2 Hoja3

Listo NUM

2.2. Una multinacional tiene un total de 40.000 trabajadores distribuidos en 400 fábricas de 100 obreros cada una. Una muestra aleatoria con probabilidades iguales sin reposición de 25 fábricas presenta la siguiente distribución de obreros mayores de 50 años:

Total de obreros mayores de 50 *años*	12	17	23	33	36
N° de fábricas de la muestra	2	3	9	5	6

Estimar el total y la proporción de obreros de la multinacional con más de 50 años, así como sus errores de muestreo absolutos y relativos.



Como el muestreo es con probabilidades iguales y se seleccionan 25 fábricas de entre 400, se tiene $\pi_i = 25/400 = 0{,}0625$ y $\pi_{ij} = (25.24)/(400.399) = 0{,}00376$. Como el método es sin reposición, tomamos como estimador del total de clase el estimador de Horwitz y Thompson y tenemos:

$$\hat{A}_{HT} = \sum_{i=1}^{25} \frac{A_i}{\pi_i} = \frac{2 \cdot 12 + 3 \cdot 17 + 9 \cdot 23 + 5 \cdot 33 + 6 \cdot 36}{25/400} = 10608$$

Para estimar la varianza tomamos el estimador de Yates y Grundy. Tenemos:

$$\hat{V}(\hat{A}_{HT}) = \sum_{i<j}^{25} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{A_i}{\pi_i} - \frac{A_j}{\pi_j} \right)^2 = \frac{0{,}0625^2 - 0{,}00376}{0{,}00376 \cdot 0{,}0625^2} \sum_{i<j}^{25} \left(A_i - A_j \right)^2 =$$

$$= 9{,}957 \left[2 \cdot 3(12-17)^2 + \cdots + 5 \cdot 6(33-36)^2 \right] = 386906{,}5$$

Las operaciones anteriores totalmente desarrolladas se muestran a continuación.

$\hat{V}(\hat{A}_{HT})$ = 9,957[2*3(12-17)^2+2*9(12-23)^2+2*5(12-33)^2+2*6(12-36)^2+3*9(17-23)^2+3*5(17-33)^2 +3*6(17-36)^2 +9*5(23-33)^2+9*6(23-36)^2+5*6(33-36)^2 = 386906,553

El error absoluto de muestreo será $\sigma(\hat{A}_{HT}) = \sqrt{386906{,}5} = 622$, con lo que el error relativo valdrá $\frac{\sigma(\hat{A}_{HT})}{\hat{A}_{HT}} \cdot 100 = \frac{622}{10608} \cdot 100 = 0{,}05863 \rightarrow 5{,}8\%$

Como estimador de la proporción de trabajadores mayores de 40 años tenemos:

$$\hat{P}_{HT} = \frac{\hat{A}_{HT}}{M} = \frac{10608}{40000} = 0{,}2642 = 26{,}42\%$$

El estimador insesgado de su varianza será :

$$\hat{V}(\hat{P}_{HT}) = \frac{\hat{V}(\hat{A}_{HT})}{M^2} = \frac{386906{,}5}{40000^2} = 0{,}000242$$

El error absoluto de muestreo será $\sigma(\hat{P}_{HT}) = \sqrt{0{,}000242} = 0{,}0155$, con lo que el error relativo valdrá $\frac{\sigma(\hat{P}_{HT})}{\hat{P}_{HT}} \cdot 100 = \frac{0{,}0155}{0{,}2642} \cdot 100 = 0{,}05863 \rightarrow 5{,}8\%$.

2.3. Considérese la población de los grupos de la materia Introducción a la Estadística que se imparte en cierta universidad. La universidad tiene 647 estudiantes de esta materia repartidos en 15 grupos con M_i estudiantes cada grupo según la tabla siguiente:

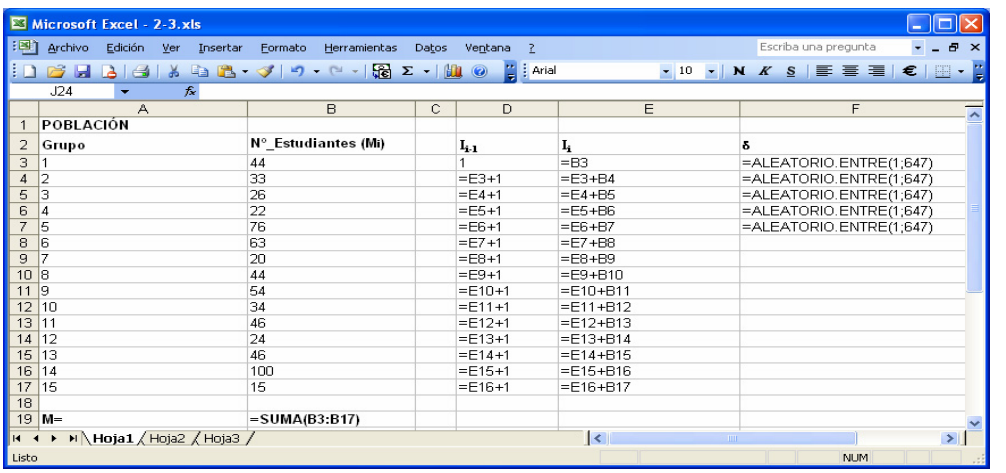
Grupos (Población)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
M_i	44	33	26	22	76	63	20	44	54	34	46	24	46	100	15

Extraemos una muestra de cinco grupos con probabilidades proporcionales a los tamaños de los grupos con reemplazo y anotamos el total de horas durante una semana que todos los estudiantes de cada grupo han empleado para estudiar la materia de Introducción a la Estadística. Los datos se recogen en la siguiente tabla:

Grupos (Muestra)	*a*	*b*	*c*	*d*	*e*
Horas	120	203	100	90	40

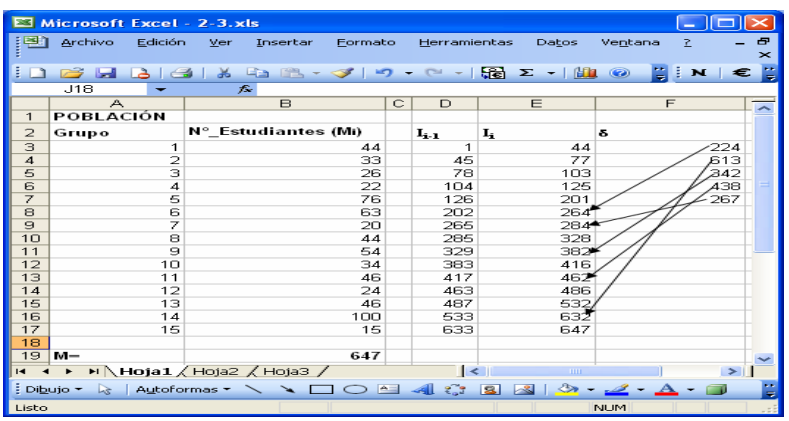
Si se ha selecciona la muestra {a, b, c, d, e} por el método del tamaño acumulativo, estimar la cantidad promedio de tiempo semanal que un estudiante empleó para estudiar la materia Introducción a la Estadística midiendo la calidad de la estimación. Estimar por intervalos al 95%.

Podemos realizar la disposición de los cálculos del método del tamaño acumulativo y la obtención de los cinco números aleatorios, como se indica en la siguiente hoja de Excel.



	A	B	C	D	E	F
1	POBLACIÓN					
2	Grupo	N°_Estudiantes (Mi)		I_{i-1}	I_i	δ
3	1	44		1	=B3	=ALEATORIO.ENTRE(1;647)
4	2	33		=E3+1	=E3+B4	=ALEATORIO.ENTRE(1;647)
5	3	26		=E4+1	=E4+B5	=ALEATORIO.ENTRE(1;647)
6	4	22		=E5+1	=E5+B6	=ALEATORIO.ENTRE(1;647)
7	5	76		=E6+1	=E6+B7	=ALEATORIO.ENTRE(1;647)
8	6	63		=E7+1	=E7+B8	
9	7	20		=E8+1	=E8+B9	
10	8	44		=E9+1	=E9+B10	
11	9	54		=E10+1	=E10+B11	
12	10	34		=E11+1	=E11+B12	
13	11	46		=E12+1	=E12+B13	
14	12	24		=E13+1	=E13+B14	
15	13	46		=E14+1	=E14+B15	
16	14	100		=E15+1	=E15+B16	
17	15	15		=E16+1	=E16+B17	
18						
19	M=	=SUMA(B3:B17)				

Realizados los cálculos, tenemos la siguiente tabla:



	A	B	C	D	E	F
1	POBLACIÓN					
2	Grupo	N°_Estudiantes (Mi)		I_{i-1}	I_i	δ
3	1	44		1	44	224
4	2	33		45	77	613
5	3	26		78	103	342
6	4	22		104	125	438
7	5	76		126	201	267
8	6	63		202	264	
9	7	20		265	284	
10	8	44		285	328	
11	9	54		329	382	
12	10	34		383	416	
13	11	46		417	462	
14	12	24		463	486	
15	13	46		487	532	
16	14	100		533	632	
17	15	15		633	647	
18						
19	M=	647				

La muestra estará formada por los grupos {6, 14, 11, 9, 7} cuyos tamaños son los siguientes:

Grupos (*Muestra*)	6	14	11	9	7
Tamaños (M_i)	63	100	54	46	20

y el número total de horas semanales empleadas por los estudiantes de esos grupos para estudiar la materia Introducción a la Estadística es el siguiente:

Grupos (*Muestra*)	6	14	11	9	7
Horas (X_i)	120	203	100	90	40

A continuación se realiza la estimación del promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística utilizando el estimador de Hansen y Hurwitz (ya que el método de selección de la muestra es con reposición). Se tiene:

$$\hat{\bar{X}}_{HH} = \frac{1}{M}\hat{X}_{HH} = \frac{1}{M}\sum_i^n \frac{X_i}{nP_i} = \frac{1}{M}\sum_i^n \frac{X_i}{n\frac{M_i}{M}} = \frac{1}{n}\sum_i^n \frac{X_i}{M_i} = \frac{1}{n}\sum_i^n \bar{X}_i = \frac{1}{5}\left(\frac{120}{63}+\frac{203}{100}+\frac{100}{54}+\frac{90}{46}+\frac{40}{20}\right) = 1{,}94$$

Por lo tanto, se estima que el promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística es algo inferior a dos horas. A continuación hallamos el error de esta estimación.

$$\hat{V}(\hat{\bar{X}}_{HH}) = \frac{1}{M^2}\hat{V}(\hat{X}_{HH}) = \frac{1}{M^2}\frac{1}{n(n-1)}\sum_{i=1}^n\left(\frac{X_i}{P_i} - \hat{X}_{HH}\right)^2 = \frac{1}{M^2}\frac{1}{n(n-1)}\sum_{i=1}^n\left(\frac{X_i}{\frac{M_i}{M}} - M\hat{\bar{X}}_{HH}\right)^2$$

$$= \frac{1}{M^2}\frac{1}{n(n-1)}\sum_{i=1}^n\left(M\frac{X_i}{M_i} - M\hat{\bar{X}}_{HH}\right)^2 = \frac{M^2}{M^2}\frac{1}{n(n-1)}\sum_{i=1}^n\left(\frac{X_i}{M_i} - \hat{\bar{X}}_{HH}\right)^2 = \frac{1}{n(n-1)}\left(\sum_{i=1}^n \bar{X}_i - \hat{\bar{X}}_{HH}\right)^2$$

con lo que se tiene:

$$\hat{V}\left(\hat{\bar{X}}_{HH}\right) = \frac{1}{5\cdot 4}\left[\left(\frac{120}{63}-1{,}94\right)^2+\left(\frac{203}{100}-1{,}94\right)^2+\left(\frac{100}{54}-1{,}94\right)^2+\left(\frac{90}{46}-1{,}94\right)^2+\left(\frac{40}{20}-1{,}94\right)^2\right] = 0{,}0034$$

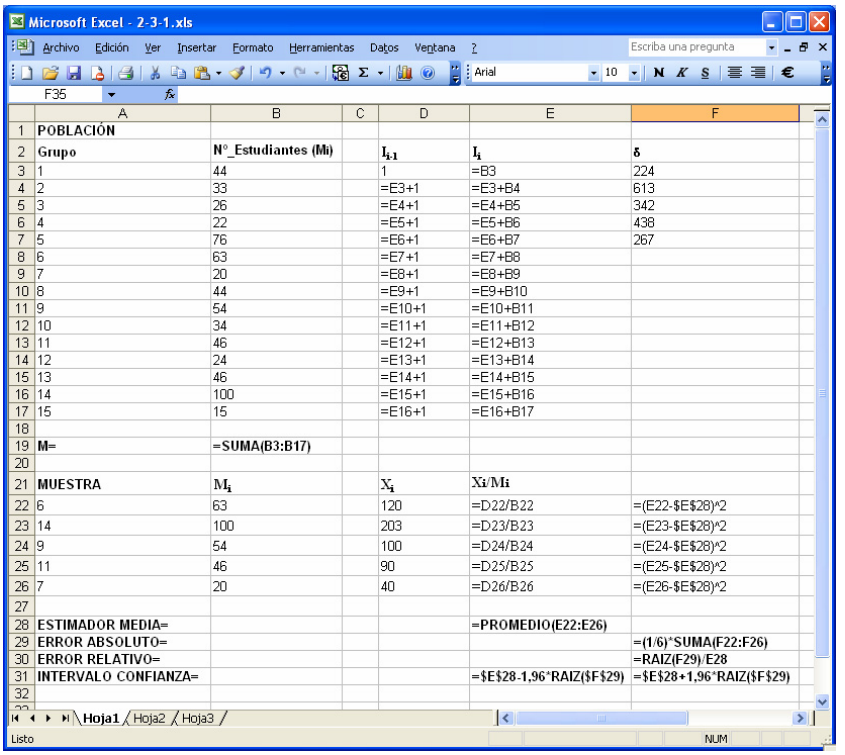
$$\hat{C}_v\left(\hat{P}\right) = \frac{\sqrt{\hat{V}\left(\hat{\bar{X}}_{HH}\right)}}{\hat{\bar{X}}_{HH}} = \frac{\sqrt{0{,}0034}}{1{,}94} = 0{,}03 \rightarrow 3\%$$

Se observa que el error relativo de muestreo es del 3%. A continuación se realiza una estimación por intervalos al 95% de confianza.

$$\hat{\bar{X}}_{HH} \pm \lambda_\alpha\sqrt{\hat{V}(\hat{\bar{X}}_{HH})} = 0{,}51 \pm 1{,}96\sqrt{0{,}0034} = [1.83,\ 2.06] \rightarrow 95\%\ \textit{confianza}$$

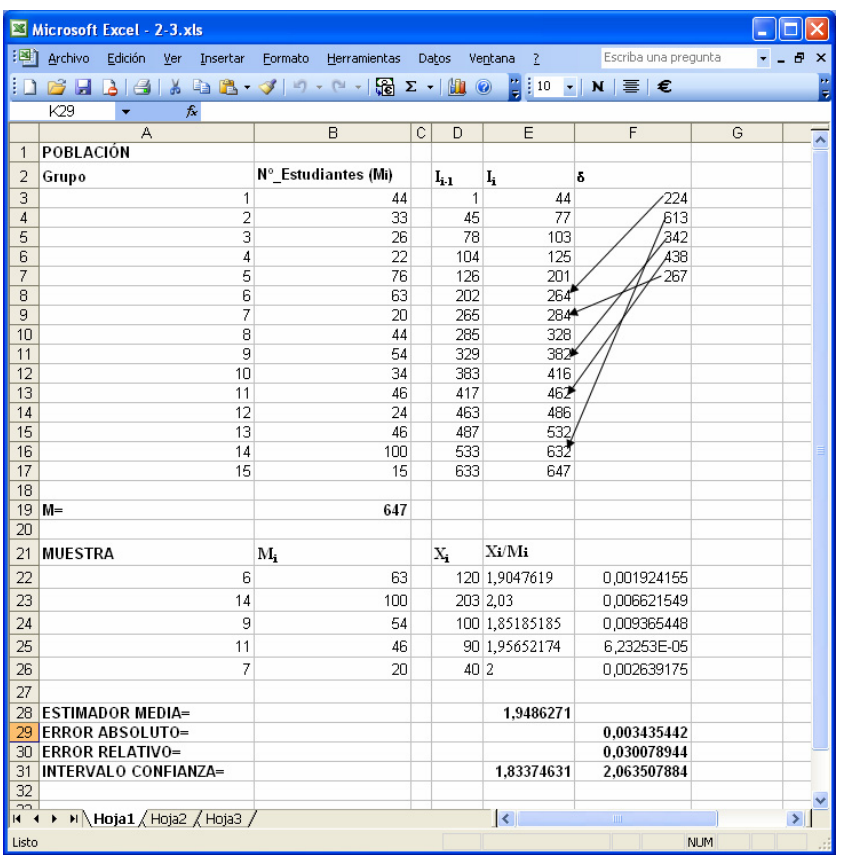
Se observa que el intervalo de confianza es muy estrecho. Esto se debe a que la estimación realizada es bastante precisa (solamente un 3% de error).

A continuación se presentan cálculos y resultados automatizados con Excel.



Microsoft Excel - 2-3-1.xls

	A	B	C	D	E	F
1	POBLACIÓN					
2	Grupo	N°_Estudiantes (Mi)		I_{i-1}	I_i	δ
3	1	44		1	=B3	224
4	2	33		=E3+1	=E3+B4	613
5	3	26		=E4+1	=E4+B5	342
6	4	22		=E5+1	=E5+B6	438
7	5	76		=E6+1	=E6+B7	267
8	6	63		=E7+1	=E7+B8	
9	7	20		=E8+1	=E8+B9	
10	8	44		=E9+1	=E9+B10	
11	9	54		=E10+1	=E10+B11	
12	10	34		=E11+1	=E11+B12	
13	11	46		=E12+1	=E12+B13	
14	12	24		=E13+1	=E13+B14	
15	13	46		=E14+1	=E14+B15	
16	14	100		=E15+1	=E15+B16	
17	15	15		=E16+1	=E16+B17	
18						
19	M=	=SUMA(B3:B17)				
20						
21	MUESTRA	M_i		X_i	Xi/Mi	
22	6	63		120	=D22/B22	=(E22-$E$28)^2
23	14	100		203	=D23/B23	=(E23-$E$28)^2
24	9	54		100	=D24/B24	=(E24-$E$28)^2
25	11	46		90	=D25/B25	=(E25-$E$28)^2
26	7	20		40	=D26/B26	=(E26-$E$28)^2
27						
28	ESTIMADOR MEDIA=				=PROMEDIO(E22:E26)	
29	ERROR ABSOLUTO=					=(1/6)*SUMA(F22:F26)
30	ERROR RELATIVO=					=RAIZ(F29)/E28
31	INTERVALO CONFIANZA=				=$E$28-1,96*RAIZ($F$29)	=$E$28+1,96*RAIZ($F$29)
32						



Microsoft Excel - 2-3.xls

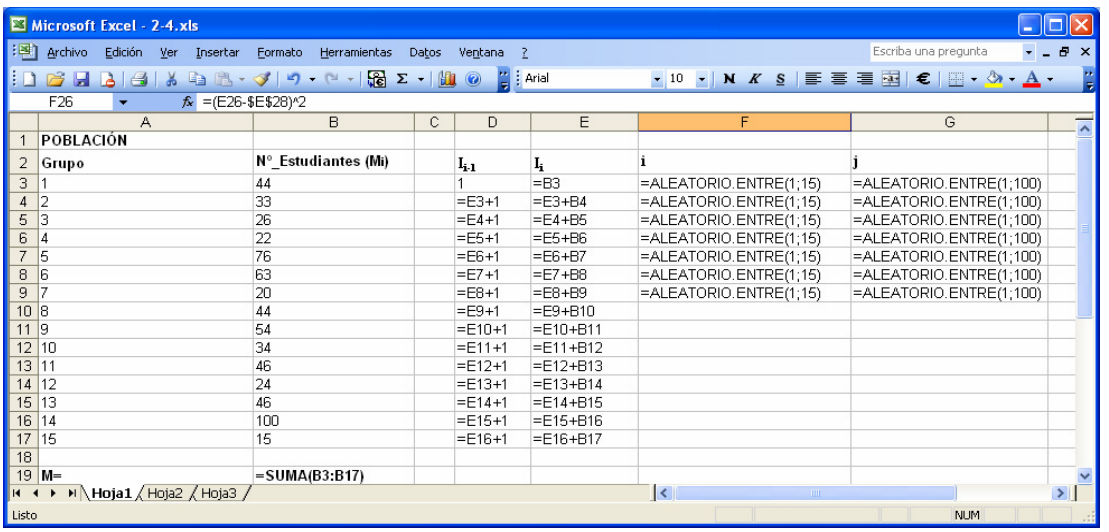
	A	B	C	D	E	F	G
1	POBLACIÓN						
2	Grupo	N°_Estudiantes (Mi)		I_{i-1}	I_i	δ	
3	1	44		1	44	224	
4	2	33		45	77	613	
5	3	26		78	103	342	
6	4	22		104	125	438	
7	5	76		126	201	267	
8	6	63		202	264		
9	7	20		265	284		
10	8	44		285	328		
11	9	54		329	382		
12	10	34		383	416		
13	11	46		417	462		
14	12	24		463	486		
15	13	46		487	532		
16	14	100		533	632		
17	15	15		633	647		
18							
19	M=	647					
20							
21	MUESTRA	M_i		X_i	Xi/Mi		
22	6	63		120	1,9047619	0,001924155	
23	14	100		203	2,03	0,006621549	
24	9	54		100	1,85185185	0,009365448	
25	11	46		90	1,95652174	6,23253E-05	
26	7	20		40	2	0,002639175	
27							
28	ESTIMADOR MEDIA=				1,9486271		
29	ERROR ABSOLUTO=					0,003435442	
30	ERROR RELATIVO=					0,030078944	
31	INTERVALO CONFIANZA=				1,83374631	2,063507884	
32							

2.4. Resolver el problema anterior suponiendo que se selecciona la muestra {a, b, c, d, e} utilizando el método de Lahiri.

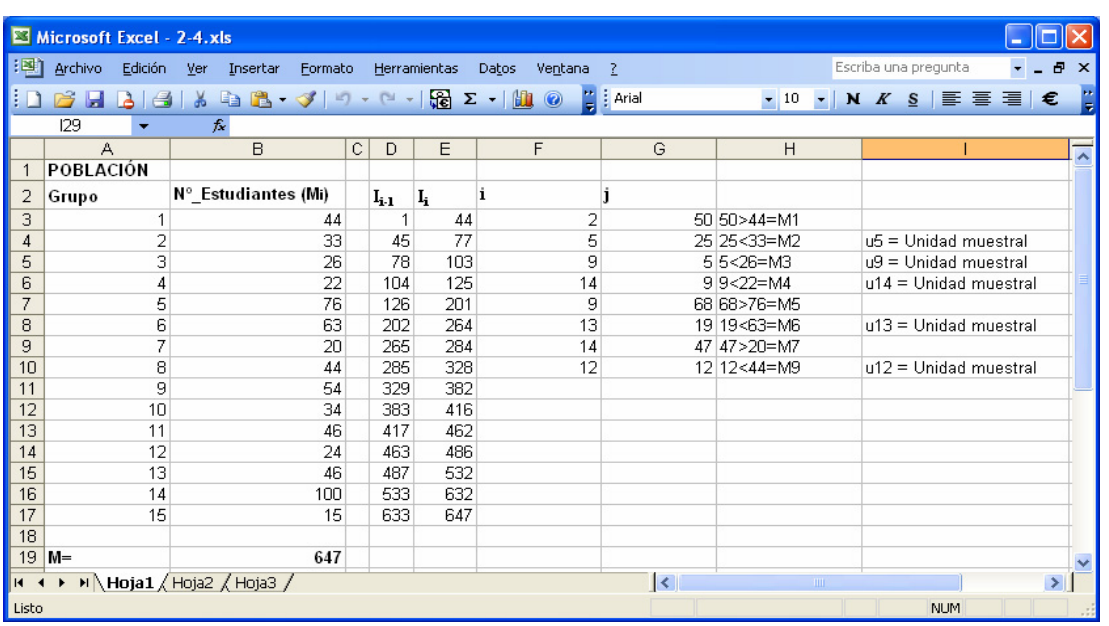
Para seleccionar la muestra mediante el método de Lahiri elegimos un par de números aleatorios (i, j) tales que $1 \le i \le N = 15$ y $1 \le j \le M_0 = \underset{i=1,2,\ldots,N}{Max}(M_i) = 100$.

Si $j \le M_i$ la unidad seleccionada para la muestra es la u_i. Si $j > M_i$, se repite la selección del par de números aleatorios (i,j) tales que $1 \le i \le N$ y $1 \le j \le M_0$ tantas veces como sea necesario hasta que $j \le M_i$.

La obtención de los números aleatorios (i,j) puede realizarse con Excel mediante la función ALEATORIO.ENTRE (que se evaluará las veces necesarias para cubrir el tamaño muestral) tal y como se indica en la figura siguiente.



El resultado obtenido es el siguiente:



La muestra estará formada por los grupos {5, 9, 14, 13, 12} cuyos tamaños son los siguientes:

Grupos (*Muestra*)	5	9	14	13	12
Tamaños (M_i)	76	54	100	46	24

y el número total de horas semanales empleadas por los estudiantes de esos grupos para estudiar la materia Introducción a la Estadística es el siguiente:

Grupos (*Muestra*)	5	9	14	13	12
Horas (X_i)	120	203	100	90	40

A continuación se realiza la estimación del promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística utilizando el estimador de Hansen y Hurwitz (ya que el método de selección de la muestra es con reposición). Se tiene:

$$\hat{\bar{X}}_{HH} = \frac{1}{M}\hat{X}_{HH} = \frac{1}{M}\sum_i^n \frac{X_i}{nP_i} = \frac{1}{M}\sum_i^n \frac{X_i}{n\frac{M_i}{M}} = \frac{1}{n}\sum_i^n \frac{X_i}{M_i} = \frac{1}{n}\sum_i^n \bar{X}_i = \frac{1}{5}\left(\frac{120}{76}+\frac{203}{54}+\frac{100}{100}+\frac{90}{46}+\frac{40}{24}\right) = 1{,}99$$

Por lo tanto, se estima que el promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística es prácticamente dos horas. A continuación hallamos el error de esta estimación.

$$\hat{V}(\hat{\bar{X}}_{HH}) = \frac{1}{M^2}\hat{V}(\hat{X}_{HH})\hat{V}(\hat{\theta}_{HH}) = \frac{1}{M^2}\frac{1}{n(n-1)}\sum_{i=1}^n\left(\frac{X_i}{P_i}-\hat{X}_{HH}\right)^2 = \frac{1}{M^2}\frac{1}{n(n-1)}\sum_{i=1}^n\left(\frac{X_i}{\frac{M_i}{M}}-M\hat{\bar{X}}_{HH}\right)^2$$

$$= \frac{1}{M^2}\frac{1}{n(n-1)}\sum_{i=1}^n\left(M\frac{X_i}{M_i}-M\hat{\bar{X}}_{HH}\right)^2 = \frac{M^2}{M^2}\frac{1}{n(n-1)}\sum_{i=1}^n\left(\frac{X_i}{M_i}-\hat{\bar{X}}_{HH}\right)^2 = \frac{1}{n(n-1)}\left(\sum_{i=1}^n \bar{X}_i - \hat{\bar{X}}_{HH}\right)^2$$

con lo que se tiene:

$$\hat{V}\left(\hat{\bar{X}}_{HH}\right) = \frac{1}{5\cdot 4}\left[\left(\frac{120}{76}-1{,}99\right)^2+\left(\frac{203}{54}-1{,}99\right)^2+\left(\frac{100}{100}-1{,}99\right)^2+\left(\frac{90}{46}-1{,}99\right)^2+\left(\frac{40}{24}-1{,}96\right)^2\right] = 0{,}73$$

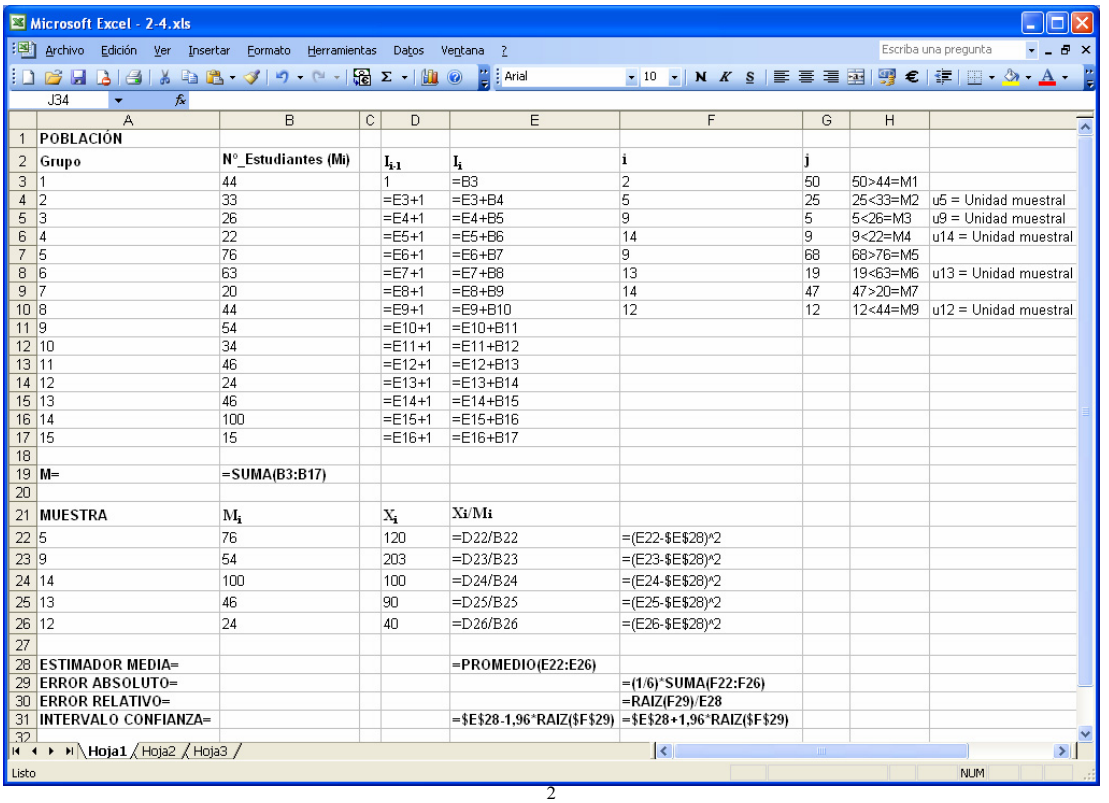
$$\hat{C}_v\left(\hat{\bar{X}}_{HH}\right) = \frac{\sqrt{\hat{V}\left(\hat{\bar{X}}_{HH}\right)}}{\hat{\bar{X}}_{HH}} = \frac{\sqrt{0{,}73}}{1{,}99} = 0{,}429 \rightarrow 43\%$$

Se observa que el error relativo de muestreo es del 43%. A continuación se realiza una estimación por intervalos al 95% de confianza.

$$\hat{\bar{X}}_{HH} \pm \lambda_\alpha\sqrt{\hat{V}(\hat{\bar{X}}_{HH})} = 1{,}99 \pm 1{,}96\sqrt{0{,}73} = [0.31,\ 3.66] \rightarrow 95\%\ \textit{confianza}$$

Se observa que el intervalo de confianza es más ancho que en el problema anterior. Esto se debe a que la estimación realizada es menos precisa (43% de error).

A continuación se presentan cálculos y resultados automatizados con Excel.



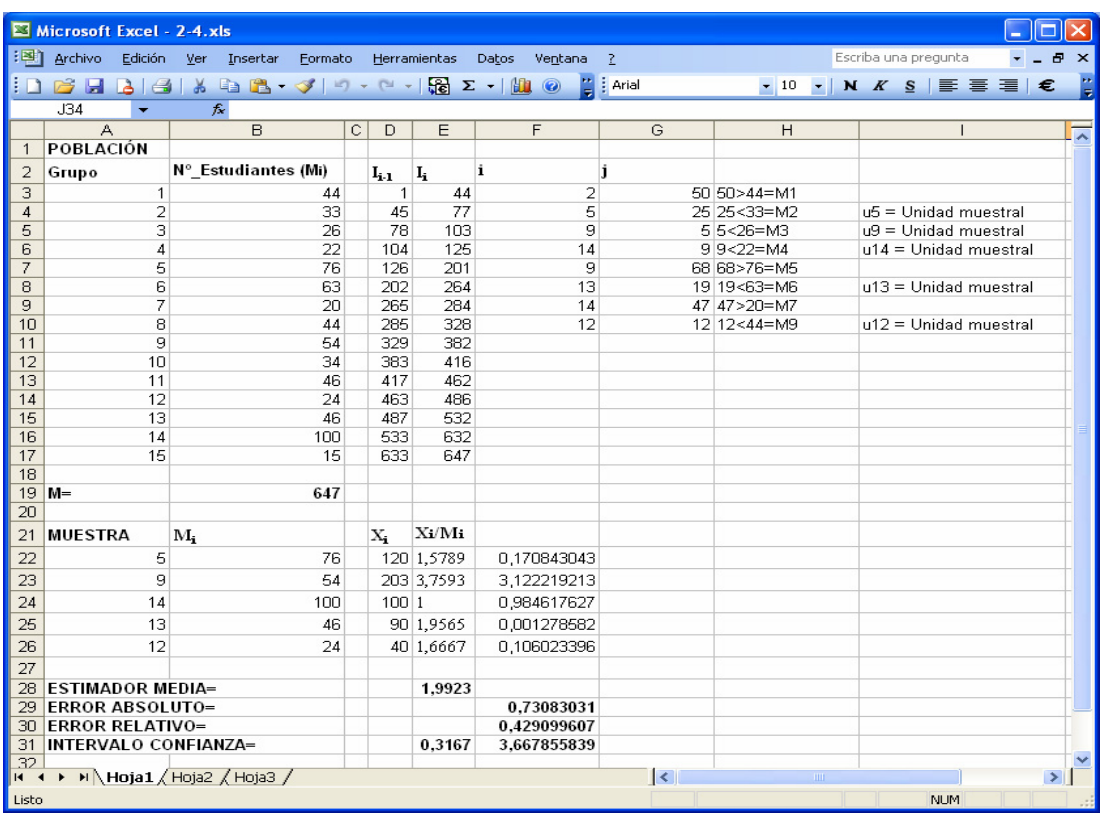
Microsoft Excel - 2-4.xls

	A	B	C	D	E	F	G	H	
1	POBLACIÓN								
2	Grupo	N°_Estudiantes (Mi)		I_{i-1}	I_i	i	j		
3	1	44		1	=B3	2	50	50>44=M1	
4	2	33		=E3+1	=E3+B4	5	25	25<33=M2	u5 = Unidad muestral
5	3	26		=E4+1	=E4+B5	9	5	5<26=M3	u9 = Unidad muestral
6	4	22		=E5+1	=E5+B6	14	9	9<22=M4	u14 = Unidad muestral
7	5	76		=E6+1	=E6+B7	9	68	68>76=M5	
8	6	63		=E7+1	=E7+B8	13	19	19<63=M6	u13 = Unidad muestral
9	7	20		=E8+1	=E8+B9	14	47	47>20=M7	
10	8	44		=E9+1	=E9+B10	12	12	12<44=M9	u12 = Unidad muestral
11	9	54		=E10+1	=E10+B11				
12	10	34		=E11+1	=E11+B12				
13	11	46		=E12+1	=E12+B13				
14	12	24		=E13+1	=E13+B14				
15	13	46		=E14+1	=E14+B15				
16	14	100		=E15+1	=E15+B16				
17	15	15		=E16+1	=E16+B17				
18									
19	M=	=SUMA(B3:B17)							
20									
21	MUESTRA	M_i		X_i	Xi/Mi				
22	5	76		120	=D22/B22	=(E22-$E$28)^2			
23	9	54		203	=D23/B23	=(E23-$E$28)^2			
24	14	100		100	=D24/B24	=(E24-$E$28)^2			
25	13	46		90	=D25/B25	=(E25-$E$28)^2			
26	12	24		40	=D26/B26	=(E26-$E$28)^2			
27									
28	ESTIMADOR MEDIA=				=PROMEDIO(E22:E26)				
29	ERROR ABSOLUTO=					=(1/6)*SUMA(F22:F26)			
30	ERROR RELATIVO=					=RAIZ(F29)/E28			
31	INTERVALO CONFIANZA=				=$E$28-1,96*RAIZ($F$29)	=$E$28+1,96*RAIZ($F$29)			
32									

Hoja1 / Hoja2 / Hoja3

Listo NUM

2



Microsoft Excel - 2-4.xls

	A	B	C	D	E	F	G	H	I
1	POBLACIÓN								
2	Grupo	N°_Estudiantes (Mi)		I_{i-1}	I_i	i	j		
3	1	44		1	44	2	50	50>44=M1	
4	2	33		45	77	5	25	25<33=M2	u5 = Unidad muestral
5	3	26		78	103	9	5	5<26=M3	u9 = Unidad muestral
6	4	22		104	125	14	9	9<22=M4	u14 = Unidad muestral
7	5	76		126	201	9	68	68>76=M5	
8	6	63		202	264	13	19	19<63=M6	u13 = Unidad muestral
9	7	20		265	284	14	47	47>20=M7	
10	8	44		285	328	12	12	12<44=M9	u12 = Unidad muestral
11	9	54		329	382				
12	10	34		383	416				
13	11	46		417	462				
14	12	24		463	486				
15	13	46		487	532				
16	14	100		533	632				
17	15	15		633	647				
18									
19	M=	647							
20									
21	MUESTRA	M_i		X_i	Xi/Mi				
22	5	76		120	1,5789	0,170843043			
23	9	54		203	3,7593	3,122219213			
24	14	100		100	1	0,984617627			
25	13	46		90	1,9565	0,001278582			
26	12	24		40	1,6667	0,106023396			
27									
28	ESTIMADOR MEDIA=				1,9923				
29	ERROR ABSOLUTO=					0,73083031			
30	ERROR RELATIVO=					0,429099607			
31	INTERVALO CONFIANZA=				0,3167	3,667855839			
32									

Hoja1 / Hoja2 / Hoja3

Listo NUM

2.5. Resolver el problema anterior suponiendo que se selecciona la muestra {a, b, c, d, e} sin reposición utilizando el método de Ikeda.

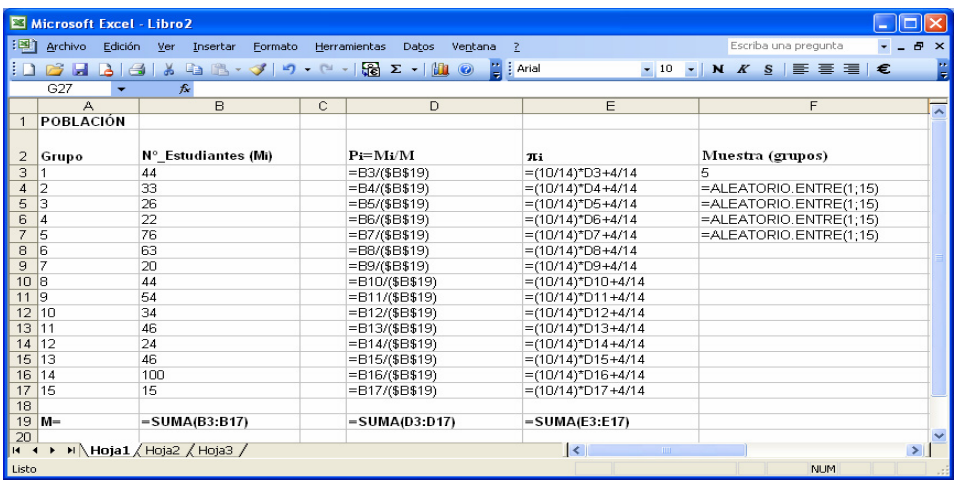
Mediante el método de Ikeda la primera unidad se obtiene sin reposición con probabilidad P_i proporcional a su tamaño M_i y las $n-1=4$ unidades restantes de la muestra se seleccionan sin reposición y con probabilidades iguales (1/4) descartando el elemento elegido inicialmente.

Los valores de π_i y π_{ij} para este método son:

$$\pi_i = \frac{N-n}{N-1} * P_i + \frac{n-1}{N-1} \qquad \pi_{ij} = \frac{n-1}{N-1} * \left[\frac{N-n}{N-2}\left(P_i + P_j\right) + \frac{n-2}{N-2} \right]$$

Para elegir la primera unidad proporcional a su tamaño podemos utilizar el método de Lahiri del problema anterior, resultando seleccionada como primera unidad muestral u_5. A continuación elegimos cuatro números aleatorios entre 1 y 15 (sin tener en cuenta el 5).

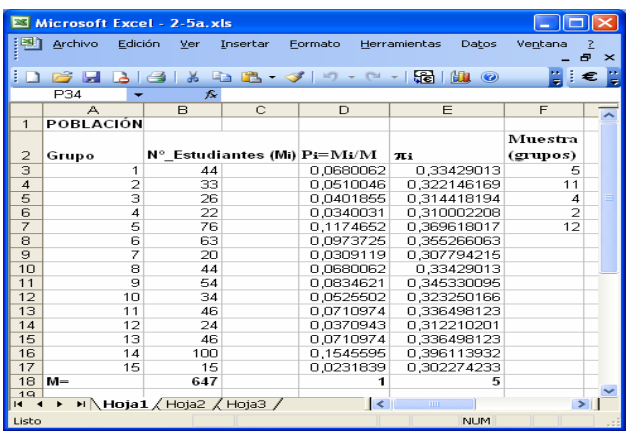
Las probabilidades P_i, π_i y los cuatro números aleatorios restantes para seleccionar las cuatro unidades que faltan para completar la muestra, pueden obtenerse como se indica en la tabla Excel siguiente.



Microsoft Excel - Libro2

	A	B	C	D	E	F
1	POBLACIÓN					
2	Grupo	N°_Estudiantes (Mi)		Pi=Mi/M	πi	Muestra (grupos)
3	1	44		=B3/($B$19)	=(10/14)*D3+4/14	5
4	2	33		=B4/($B$19)	=(10/14)*D4+4/14	=ALEATORIO.ENTRE(1;15)
5	3	26		=B5/($B$19)	=(10/14)*D5+4/14	=ALEATORIO.ENTRE(1;15)
6	4	22		=B6/($B$19)	=(10/14)*D6+4/14	=ALEATORIO.ENTRE(1;15)
7	5	76		=B7/($B$19)	=(10/14)*D7+4/14	=ALEATORIO.ENTRE(1;15)
8	6	63		=B8/($B$19)	=(10/14)*D8+4/14	
9	7	20		=B9/($B$19)	=(10/14)*D9+4/14	
10	8	44		=B10/($B$19)	=(10/14)*D10+4/14	
11	9	54		=B11/($B$19)	=(10/14)*D11+4/14	
12	10	34		=B12/($B$19)	=(10/14)*D12+4/14	
13	11	46		=B13/($B$19)	=(10/14)*D13+4/14	
14	12	24		=B14/($B$19)	=(10/14)*D14+4/14	
15	13	46		=B15/($B$19)	=(10/14)*D15+4/14	
16	14	100		=B16/($B$19)	=(10/14)*D16+4/14	
17	15	15		=B17/($B$19)	=(10/14)*D17+4/14	
18						
19	M=	=SUMA(B3:B17)		=SUMA(D3:D17)	=SUMA(E3:E17)	
20						

El resultado obtenido es el siguiente:



Microsoft Excel - 2-5a.xls

	A	B	C	D	E	F
1	POBLACIÓN					
2	Grupo	N°_Estudiantes (Mi)		Pi=Mi/M	πi	Muestra (grupos)
3	1	44		0,0680062	0,33429013	5
4	2	33		0,0510046	0,322146169	11
5	3	26		0,0401855	0,314418194	4
6	4	22		0,0340031	0,310002208	2
7	5	76		0,1174652	0,369618017	12
8	6	63		0,0973725	0,355266063	
9	7	20		0,0309119	0,307794215	
10	8	44		0,0680062	0,33429013	
11	9	54		0,0834621	0,345330095	
12	10	34		0,0525502	0,323250166	
13	11	46		0,0710974	0,336498123	
14	12	24		0,0370943	0,312210201	
15	13	46		0,0710974	0,336498123	
16	14	100		0,1545595	0,396113932	
17	15	15		0,0231839	0,302274233	
18	M=	647		1	5	

La muestra estará formada por los grupos {5, 9, 14, 13, 12} cuyos tamaños son los siguientes:

Grupos (*Muestra*)	5	11	4	2	12
Tamaños (M_i)	76	46	22	33	24

y el número total de horas semanales empleadas por los estudiantes de esos grupos para estudiar la materia Introducción a la Estadística es el siguiente:

Grupos (*Muestra*)	5	11	4	2	12
Horas (X_i)	120	203	100	90	40

A continuación se realiza la estimación del promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística utilizando el estimador de Horvitz y Thompson (ya que el método de selección de la muestra es sin reposición). En la tabla siguiente se presentan todos los cálculos necesarios para realizar la estimación (N=15, n=5).

MUESTRA	**M_i**	**Pi=Mi/M**	**πi = Pi (N-n)/(N-1) + (n-1)/(N-1)**	**Xi**	**Xi/πi**
5	76	0,1174652	0,369618017	120	324,659
11	46	0,0710974	0,336498123	203	603,272
4	22	0,0340031	0,310002208	100	322,578
2	33	0,0510046	0,322146169	90	279,376
12	24	0,0370943	0,312210201	40	128,119
SUMA=					**1658,01**

$$\hat{\bar{X}}_{HT} = \frac{1}{M}\hat{X}_{HT} = \frac{1}{M}\sum_{i=1}^{25}\frac{X_i}{\pi_i} = \frac{1}{647}\left(\frac{120}{0{,}369} + \frac{203}{0{,}336} + \frac{100}{0{,}310} + \frac{90}{0{,}322} + \frac{40}{0{,}312}\right) = \frac{1658}{647} = 2{,}56$$

Por lo tanto, se estima que el promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística es prácticamente dos horas y media. A continuación hallamos el error de esta estimación a través de la estimación de la varianza. En la siguiente tabla se presentan todos los cálculos necesarios para realizar la estimación (N=15, n=5).

$(X_i/\pi_i)^2(1-\pi_i)$	**π_{ij}**	**X_i**	**X_j**	**π_i**	**π_j**	**P_i**	**P_j**	**$(X_i/\pi_i)(X_j/\pi_j)(\pi_{ij}-\pi_i\pi_j)/\pi_{ij}$**
66444,64999	0,107	120	203	0,3696	0,3365	0,117	0,0711	-31007,41923
241473,2008	0,099	120	100	0,3696	0,31	0,117	0,034	-16210,59516
71798,95557	0,103	120	90	0,3696	0,3221	0,117	0,051	-14192,52368
52907,24303	0,1	120	40	0,3696	0,3122	0,117	0,0371	-6451,525615
11289,67748	0,089	203	100	0,3365	0,31	0,071	0,034	-33402,58498
SUMA=443913,7	0,093	203	90	0,3365	0,3221	0,071	0,051	-28399,38423
	0,09	203	40	0,3365	0,3122	0,071	0,0371	-13220,87321
	0,085	100	90	0,31	0,3221	0,034	0,051	-16240,761
	0,082	100	40	0,31	0,3122	0,034	0,0371	-7715,377148
	0,085	90	40	0,3221	0,3122	0,051	0,0371	-6412,377049
								2*SUMA= -346506,8426

$$\hat{V}\left(\hat{\bar{X}}_{HT}\right)=\frac{1}{M^2}\hat{V}\left(\hat{X}_{HT}\right)=\frac{1}{M^2}\left[\sum_{i=1}^{5}\frac{X_i^2}{\pi_i^2}(1-\pi_i)+2\sum_{i=1}^{5}\sum_{j>i}^{5}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\right]=$$

$$\frac{1}{M^2}\left[\frac{X_1^2}{\pi_1}(1-\pi_1)+\cdots+\frac{X_5^2}{\pi_5}(1-\pi_5)+2\left(\frac{X_1}{\pi_1}\frac{X_2}{\pi_2}\left(\frac{\pi_{12}-\pi_1\pi_2}{\pi_{12}}\right)+\cdots+\frac{X_4}{\pi_4}\frac{X_5}{\pi_5}\left(\frac{\pi_{45}-\pi_4\pi_5}{\pi_{45}}\right)\right)\right]=$$

$$=\frac{443913{,}7269-346506{,}8426}{647^2}=0{,}232692$$

con lo que se tiene:

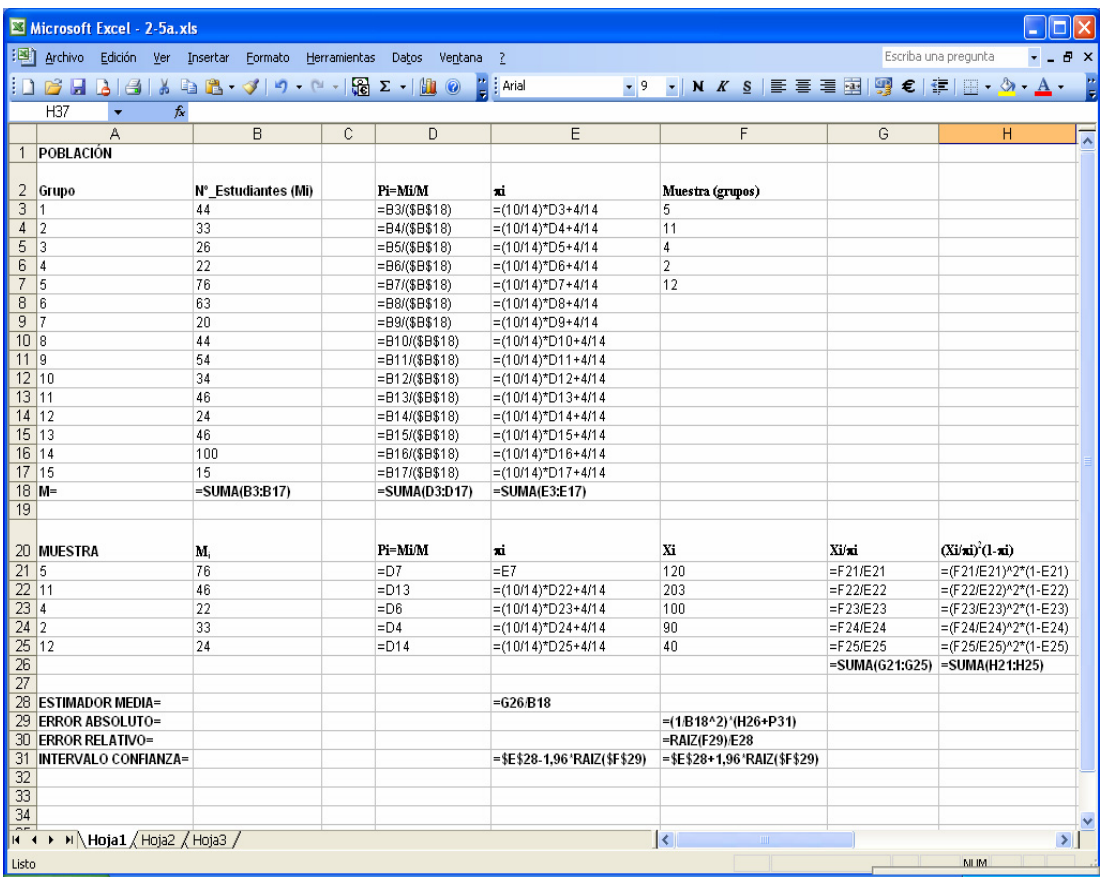
$$\hat{C}_v\left(\hat{\bar{X}}_{HT}\right)=\frac{\sqrt{\hat{V}\left(\hat{\bar{X}}_{HT}\right)}}{\hat{\bar{X}}_{HT}}=\frac{\sqrt{0{,}232692}}{2{,}56}=0{,}188\rightarrow 19\%$$

Se observa que el error relativo de muestreo es del 43%. A continuación se realiza una estimación por intervalos al 95% de confianza.

$$\hat{\bar{X}}_{HT}\pm\lambda_\alpha\sqrt{\hat{V}(\hat{\bar{X}}_{HT})}=2{,}56\pm 1{,}96\sqrt{0{,}232692}=[1.617,\ 3.508]\rightarrow 95\%\ \textit{confianza}$$

Se observa que el intervalo de confianza es más estrecho que en el problema anterior. Esto se debe a que la estimación realizada es más precisa (19% de error).

A continuación se presentan cálculos y resultados automatizados con Excel.



Microsoft Excel - 2-5a.xls

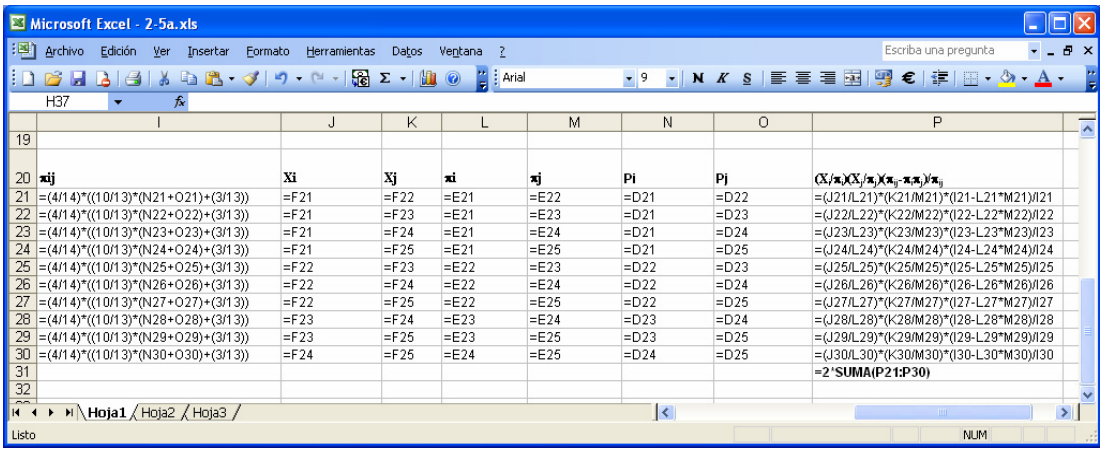
Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

H37

	A	B	C	D	E	F	G	H
1	POBLACIÓN							
2	Grupo	Nº_Estudiantes (Mi)		Pi=Mi/M	xi	Muestra (grupos)		
3	1	44		=B3/($B$18)	=(10/14)*D3+4/14	5		
4	2	33		=B4/($B$18)	=(10/14)*D4+4/14	11		
5	3	26		=B5/($B$18)	=(10/14)*D5+4/14	4		
6	4	22		=B6/($B$18)	=(10/14)*D6+4/14	2		
7	5	76		=B7/($B$18)	=(10/14)*D7+4/14	12		
8	6	63		=B8/($B$18)	=(10/14)*D8+4/14			
9	7	20		=B9/($B$18)	=(10/14)*D9+4/14			
10	8	44		=B10/($B$18)	=(10/14)*D10+4/14			
11	9	54		=B11/($B$18)	=(10/14)*D11+4/14			
12	10	34		=B12/($B$18)	=(10/14)*D12+4/14			
13	11	46		=B13/($B$18)	=(10/14)*D13+4/14			
14	12	24		=B14/($B$18)	=(10/14)*D14+4/14			
15	13	46		=B15/($B$18)	=(10/14)*D15+4/14			
16	14	100		=B16/($B$18)	=(10/14)*D16+4/14			
17	15	15		=B17/($B$18)	=(10/14)*D17+4/14			
18	M=	=SUMA(B3:B17)		=SUMA(D3:D17)	=SUMA(E3:E17)			
19								
20	MUESTRA	M_i		Pi=Mi/M	xi	Xi	Xi/xi	$(Xi/xi)^2(1-xi)$
21	5	76		=D7	=E7	120	=F21/E21	=(F21/E21)^2*(1-E21)
22	11	46		=D13	=(10/14)*D22+4/14	203	=F22/E22	=(F22/E22)^2*(1-E22)
23	4	22		=D6	=(10/14)*D23+4/14	100	=F23/E23	=(F23/E23)^2*(1-E23)
24	2	33		=D4	=(10/14)*D24+4/14	90	=F24/E24	=(F24/E24)^2*(1-E24)
25	12	24		=D14	=(10/14)*D25+4/14	40	=F25/E25	=(F25/E25)^2*(1-E25)
26							=SUMA(G21:G25)	=SUMA(H21:H25)
27								
28	ESTIMADOR MEDIA=				=G26/B18			
29	ERROR ABSOLUTO=					=(1/B18^2)*(H26+P31)		
30	ERROR RELATIVO=					=RAIZ(F29)/E28		
31	INTERVALO CONFIANZA=				=$E$28-1,96*RAIZ($F$29)	=$E$28+1,96*RAIZ($F$29)		
32								
33								
34								

Hoja1 Hoja2 Hoja3

Listo NUM



Microsoft Excel - 2-5a.xls

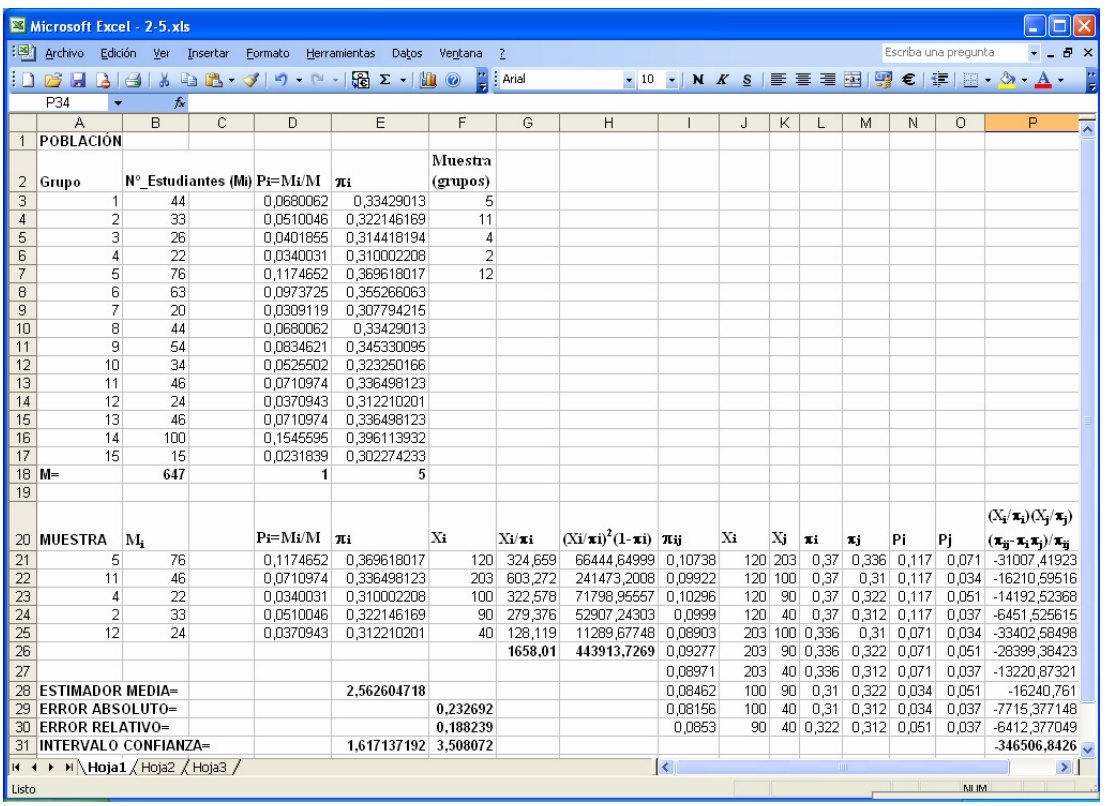
Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

H37

	I	J	K	L	M	N	O	P
19								
20	**πij**	**Xi**	**Xj**	**πi**	**πj**	**Pi**	**Pj**	**$(X_i/\pi_i)(X_j/\pi_j)(\pi_{ij}-\pi_i\pi_j)/\pi_{ij}$**
21	=(4/14)*((10/13)*(N21+O21)+(3/13))	=F21	=F22	=E21	=E22	=D21	=D22	=(J21/L21)*(K21/M21)*(I21-L21*M21)/I21
22	=(4/14)*((10/13)*(N22+O22)+(3/13))	=F21	=F23	=E21	=E23	=D21	=D23	=(J22/L22)*(K22/M22)*(I22-L22*M22)/I22
23	=(4/14)*((10/13)*(N23+O23)+(3/13))	=F21	=F24	=E21	=E24	=D21	=D24	=(J23/L23)*(K23/M23)*(I23-L23*M23)/I23
24	=(4/14)*((10/13)*(N24+O24)+(3/13))	=F21	=F25	=E21	=E25	=D21	=D25	=(J24/L24)*(K24/M24)*(I24-L24*M24)/I24
25	=(4/14)*((10/13)*(N25+O25)+(3/13))	=F22	=F23	=E22	=E23	=D22	=D23	=(J25/L25)*(K25/M25)*(I25-L25*M25)/I25
26	=(4/14)*((10/13)*(N26+O26)+(3/13))	=F22	=F24	=E22	=E24	=D22	=D24	=(J26/L26)*(K26/M26)*(I26-L26*M26)/I26
27	=(4/14)*((10/13)*(N27+O27)+(3/13))	=F22	=F25	=E22	=E25	=D22	=D25	=(J27/L27)*(K27/M27)*(I27-L27*M27)/I27
28	=(4/14)*((10/13)*(N28+O28)+(3/13))	=F23	=F24	=E23	=E24	=D23	=D24	=(J28/L28)*(K28/M28)*(I28-L28*M28)/I28
29	=(4/14)*((10/13)*(N29+O29)+(3/13))	=F23	=F25	=E23	=E25	=D23	=D25	=(J29/L29)*(K29/M29)*(I29-L29*M29)/I29
30	=(4/14)*((10/13)*(N30+O30)+(3/13))	=F24	=F25	=E24	=E25	=D24	=D25	=(J30/L30)*(K30/M30)*(I30-L30*M30)/I30
31								=2*SUMA(P21:P30)
32								

Hoja1 Hoja2 Hoja3

Listo NUM



Microsoft Excel - 2-5.xls

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

P34

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**POBLACIÓN**															
2	**Grupo**	**N°_Estudiantes (Mi)**		**Pi=Mi/M**	**πi**	**Muestra (grupos)**										
3	1	44		0,0680062	0,33429013	5										
4	2	33		0,0510046	0,322146169	11										
5	3	26		0,0401855	0,314418194	4										
6	4	22		0,0340031	0,310002208	2										
7	5	76		0,1174652	0,369618017	12										
8	6	63		0,0973725	0,355266063											
9	7	20		0,0309119	0,307794215											
10	8	44		0,0680062	0,33429013											
11	9	54		0,0834621	0,345330095											
12	10	34		0,0525502	0,323250166											
13	11	46		0,0710974	0,336498123											
14	12	24		0,0370943	0,312210201											
15	13	46		0,0710974	0,336498123											
16	14	100		0,1545595	0,396113932											
17	15	15		0,0231839	0,302274233											
18	**M=**	**647**		**1**	**5**											
19																
20	**MUESTRA**	**M_i**		**Pi=Mi/M**	**πi**	**Xi**	**Xi/πi**	**$(Xi/\pi i)^2(1-\pi i)$**	**πij**	**Xi**	**Xj**	**πi**	**πj**	**Pi**	**Pj**	**$(X_i/\pi_i)(X_j/\pi_j)$ $(\pi_{ij}-\pi_i\pi_j)/\pi_{ij}$**
21	5	76		0,1174652	0,369618017	120	324,659	66444,64999	0,10738	120	203	0,37	0,336	0,117	0,071	-31007,41923
22	11	46		0,0710974	0,336498123	203	603,272	241473,2008	0,09922	120	100	0,37	0,31	0,117	0,034	-16210,59516
23	4	22		0,0340031	0,310002208	100	322,578	71798,95557	0,10296	120	90	0,37	0,322	0,117	0,051	-14192,52368
24	2	33		0,0510046	0,322146169	90	279,376	52907,24303	0,0999	120	40	0,37	0,312	0,117	0,037	-6451,525615
25	12	24		0,0370943	0,312210201	40	128,119	11289,67748	0,08903	203	100	0,336	0,31	0,071	0,034	-33402,58498
26							**1658,01**	**443913,7269**	0,09277	203	90	0,336	0,322	0,071	0,051	-28399,38423
27									0,08971	203	40	0,336	0,312	0,071	0,037	-13220,87321
28	**ESTIMADOR MEDIA=**				**2,562604718**				0,08462	100	90	0,31	0,322	0,034	0,051	-16240,761
29	**ERROR ABSOLUTO=**					**0,232692**			0,08156	100	40	0,31	0,312	0,034	0,037	-7715,377148
30	**ERROR RELATIVO=**					**0,188239**			0,0853	90	40	0,322	0,312	0,051	0,037	-6412,377049
31	**INTERVALO CONFIANZA=**				**1,617137192**	**3,508072**										**-346506,8426**

Hoja1 Hoja2 Hoja3

Listo NUM

2.6. Resolver el problema anterior suponiendo que se selecciona la muestra {a, b, c, d, e} sin reposición utilizando el método de Sampford.

En el método de Sampford los elementos muestrales se eligen con reposición seleccionando el primer elemento con probabilidad P_i y los restantes $n - 1$ elementos con probabilidades proporcionales a $P_i/(1-nP_i)$. Finalizada la extracción, la muestra se acepta si todos los elementos son diferentes, y en caso contrario se rechaza y se vuelve a empezar.

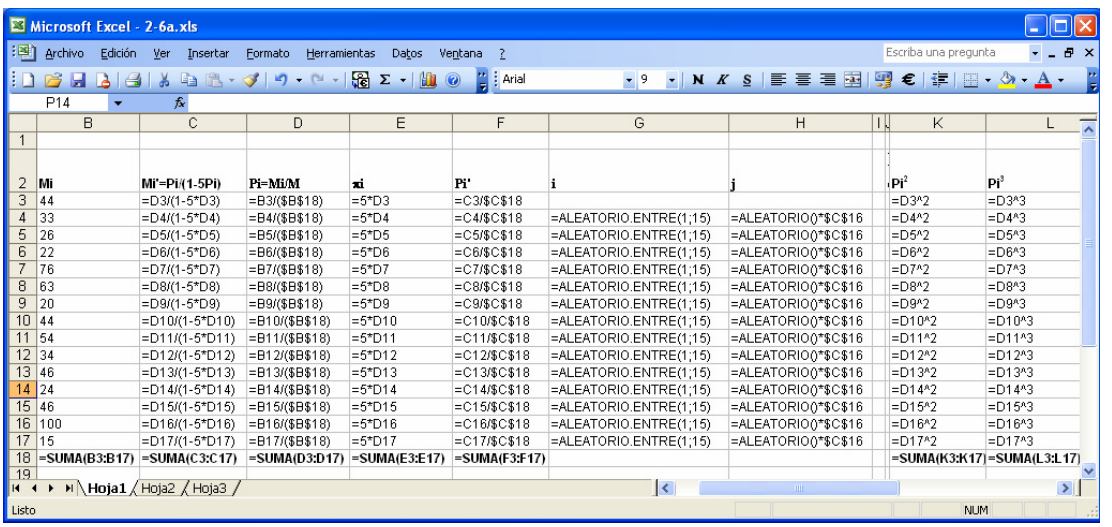
Mediante este método se tiene que:

$$\pi_i = nP_i$$

$$\pi_{ij} \approx n(n-1)P_iP_j\left(\begin{array}{l}1+\left[\left(P_i+P_j\right)-\sum_k P_k^2\right]+2\left(P_i^2+P_j^2\right)-2\sum_k P_k^3-(n-2)P_iP_j+\\ +(n-3)\left(P_i+P_j\right)-\sum_k P_k^3-(n-3)\left(\sum_k P_k^2\right)\end{array}\right)$$

Para llevar a la práctica el método de Sampford se calculan los $P_i = M_i/M$ y a continuación se hallan ΣP_i^2 y ΣP_i^3, valores que se utilizarán para calculai π_{ij}. La siguiente tarea es calcular $P_i/(1-5P_i)$, $P_i' = [P_i/(1-5P_i)]/\Sigma\,[P_i/(1-5P_i)]$ y $\pi_i = 5P_i$.

La siguiente tarea es extraer las cinco unidades muestrales. La primera unidad se extrae con probabilidad $P_i = M_i/M$ proporcional a su tamaño M_i y las siguientes unidades se extraen con probabilidades $P_i' = [P_i/(1-5P_i)]/\Sigma[P_i/(1-5P_i)]$ proporcionales a $P_i/(1-5P_i)$, con reemplazamiento. Si sale alguna unidad repetida se repiten otra vez todas las extracciones hasta que no salga ninguna repetida. Para elegir la primera unidad proporcional a su tamaño podemos utilizar el método de Lahiri del problema 2.4, resultando seleccionada como primera unidad muestral u_5. Para elegir las cuatro siguientes unidades volvemos a repetir el método de Lahiri [extracción de pares de números aleatorios (i,j) con $1 \le i \le 15$ y $0 \le j \le Máx(M_i') = 0{,}68$ hasta que $j \le M_i'$] y resultan elegidas u_3, u_8, u_{13} y u_{14}. Las tablas siguientes ilustran las fórmulas con Excel y los resultados obtenidos.



	B	C	D	E	F	G	H	K	L
1									
2	**Mi**	**Mi'=Pi/(1-5Pi)**	**Pi=Mi/M**	**πi**	**Pi'**	**i**	**j**	**Pi²**	**Pi³**
3	44	=D3/(1-5*D3)	=B3/($B$18)	=5*D3	=C3/$C$18			=D3^2	=D3^3
4	33	=D4/(1-5*D4)	=B4/($B$18)	=5*D4	=C4/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D4^2	=D4^3
5	26	=D5/(1-5*D5)	=B5/($B$18)	=5*D5	=C5/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D5^2	=D5^3
6	22	=D6/(1-5*D6)	=B6/($B$18)	=5*D6	=C6/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D6^2	=D6^3
7	76	=D7/(1-5*D7)	=B7/($B$18)	=5*D7	=C7/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D7^2	=D7^3
8	63	=D8/(1-5*D8)	=B8/($B$18)	=5*D8	=C8/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D8^2	=D8^3
9	20	=D9/(1-5*D9)	=B9/($B$18)	=5*D9	=C9/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D9^2	=D9^3
10	44	=D10/(1-5*D10)	=B10/($B$18)	=5*D10	=C10/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D10^2	=D10^3
11	54	=D11/(1-5*D11)	=B11/($B$18)	=5*D11	=C11/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D11^2	=D11^3
12	34	=D12/(1-5*D12)	=B12/($B$18)	=5*D12	=C12/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D12^2	=D12^3
13	46	=D13/(1-5*D13)	=B13/($B$18)	=5*D13	=C13/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D13^2	=D13^3
14	24	=D14/(1-5*D14)	=B14/($B$18)	=5*D14	=C14/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D14^2	=D14^3
15	46	=D15/(1-5*D15)	=B15/($B$18)	=5*D15	=C15/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D15^2	=D15^3
16	100	=D16/(1-5*D16)	=B16/($B$18)	=5*D16	=C16/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D16^2	=D16^3
17	15	=D17/(1-5*D17)	=B17/($B$18)	=5*D17	=C17/$C$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$C$16	=D17^2	=D17^3
18	=SUMA(B3:B17)	=SUMA(C3:C17)	=SUMA(D3:D17)	=SUMA(E3:E17)	=SUMA(F3:F17)			=SUMA(K3:K17)	=SUMA(L3:L17)
19									

	Mi	Pi=Mi/M	Mi'=Pi/(1-5Pi)	πi	Pi'	i	j		Ui	Pi^2	Pi^3
1	44	0,0680062	0,103044	0,34	0,0499				5	0,0046	0,00031
2	33	0,0510046	0,068465	0,255	0,0332	4	0,613			0,0026	0,00013
3	26	0,0401855	0,05029	0,2009	0,0244	14	0,037	<M'3=0,04	3	0,0016	6,5E-05
4	22	0,0340031	0,040968	0,17	0,0198	14	0,665			0,0012	3,9E-05
5	76	0,1174652	0,284644	0,5873	0,1379	4	0,365			0,0138	0,00162
6	63	0,0973725	0,189759	0,4869	0,0919	13	0,514			0,0095	0,00092
7	20	0,0309119	0,036563	0,1546	0,0177	2	0,585			0,001	3E-05
8	44	0,0680062	0,103044	0,34	0,0499	4	0,096	<M'8=0,10	8	0,0046	0,00031
9	54	0,0834621	0,143236	0,4173	0,0694	14	0,231			0,007	0,00058
10	34	0,0525502	0,071279	0,2628	0,0345	11	0,081			0,0028	0,00015
11	46	0,0710974	0,110312	0,3555	0,0534	15	0,674			0,0051	0,00036
12	24	0,0370943	0,045541	0,1855	0,0221	3	0,295			0,0014	5,1E-05
13	46	0,0710974	0,110312	0,3555	0,0534	13	0,054	<M'13=0,11	13	0,0051	0,00036
14	100	0,1545595	**0,680272**	0,7728	0,3296	1	0,319	<M'14=0,68	14	0,0239	0,00369
15	15	0,0231839	0,026224	0,1159	0,0127					0,0005	1,2E-05
SUMA	**647**	**1**	**2,063954**	**5**	**1**					**0,0845**	**0,00864**

La muestra estará formada por los grupos {5, 3, 8, 13, 14} cuyos tamaños son los siguientes:

Grupos (*Muestra*)	5	3	8	13	14
Tamaños (M_i)	76	26	44	46	100

y el número total de horas semanales empleadas por los estudiantes de esos grupos para estudiar la materia Introducción a la Estadística es el siguiente:

Grupos (*Muestra*)	5	3	8	13	14
Horas (X_i)	120	203	100	90	40

A continuación se realiza la estimación del promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística utilizando el estimador de Horvitz y Thompson (ya que el método de selección de la muestra es sin reposición). En la siguiente tabla se presentan todos los cálculos necesarios para realizar la estimación (N=15, n=5).

MUESTRA	**M_i**	**Pi=Mi/M**	**$\pi i = 5Pi$**	**Xi**	**$Xi/\pi i$**	**$(Xi/\pi i)^2(1-\pi i)$**
5	76	0,1174652	0,5873	120	204,32	17227,0471
3	26	0,0401855	0,2009	203	1010,3	815643,153
8	44	0,0680062	0,34	100	294,09	57080,3719
13	46	0,0710974	0,3555	90	253,17	41311,3781
14	100	0,1545595	0,7728	40	51,76	608,6976
					1813,7	**931870,648**

$$\hat{\bar{X}}_{HT} = \frac{1}{M}\hat{X}_{HT} = \frac{1}{M}\sum_{i=1}^{25}\frac{X_i}{\pi_i} = \frac{1}{647}\left(\frac{120}{0{,}587} + \frac{203}{0{,}201} + \frac{100}{0{,}340} + \frac{90}{0{,}355} + \frac{40}{0{,}772}\right) = \frac{1813{,}7}{647} = 2{,}8$$

Por lo tanto, se estima que el promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística es 2,8 horas. A continuación hallamos el error de esta estimación a través de la estimación de la varianza. Ahora se presenta la tabla con todos los cálculos necesarios para realizar la estimación (N = 15, n = 5).

$(X_i/\pi_i)^2(1-\pi_i)$	π_{ij}	X_i	X_j	π_i	π_j	P_i	P_j	$(X_i/\pi_i)(X_j/\pi_j)(\pi_{ij}-\pi_i\pi_j)/\pi_{ij}$
17227,0471	0,1024232	120	203	0,58733	0,201	0,117	0,04	-31413,39362
815643,153	0,177923	120	100	0,58733	0,34	0,117	0,068	-7357,500123
57080,3719	0,1865756	120	90	0,58733	0,355	0,117	0,071	-6157,939465
41311,3781	0,4440242	120	40	0,58733	0,773	0,117	0,155	-234,8366464
608,6976	0,0554606	203	100	0,20093	0,34	0,04	0,068	-68900,81915
931870,648	0,0582159	203	90	0,20093	0,355	0,04	0,071	-58046,26694
	0,1421053	203	40	0,20093	0,773	0,04	0,155	-4846,785171
	0,1016725	100	90	0,34003	0,355	0,068	0,071	-14063,38541
	0,2458765	100	40	0,34003	0,773	0,068	0,155	-1046,18541
	0,2577209	90	40	0,35549	0,773	0,071	0,155	-864,3197294
								-385862,8633

$$\hat{V}\left(\hat{\bar{X}}_{HT}\right)=\frac{1}{M^2}\hat{V}\left(\hat{X}_{HT}\right)=\frac{1}{M^2}\left[\sum_{i=1}^{5}\frac{X_i^2}{\pi_i^2}(1-\pi_i)+2\sum_{i=1}^{5}\sum_{j>i}^{5}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\right]=$$

$$\frac{1}{M^2}\left[\frac{X_1^2}{\pi_1}(1-\pi_1)+\cdots+\frac{X_5^2}{\pi_5}(1-\pi_5)+2\left(\frac{X_1}{\pi_1}\frac{X_2}{\pi_2}\left(\frac{\pi_{12}-\pi_1\pi_2}{\pi_{12}}\right)+\cdots+\frac{X_4}{\pi_4}\frac{X_5}{\pi_5}\left(\frac{\pi_{45}-\pi_4\pi_5}{\pi_{45}}\right)\right)\right]=$$

$$=\frac{931870{,}648-385862{,}8633}{647^2}=1{,}3043$$

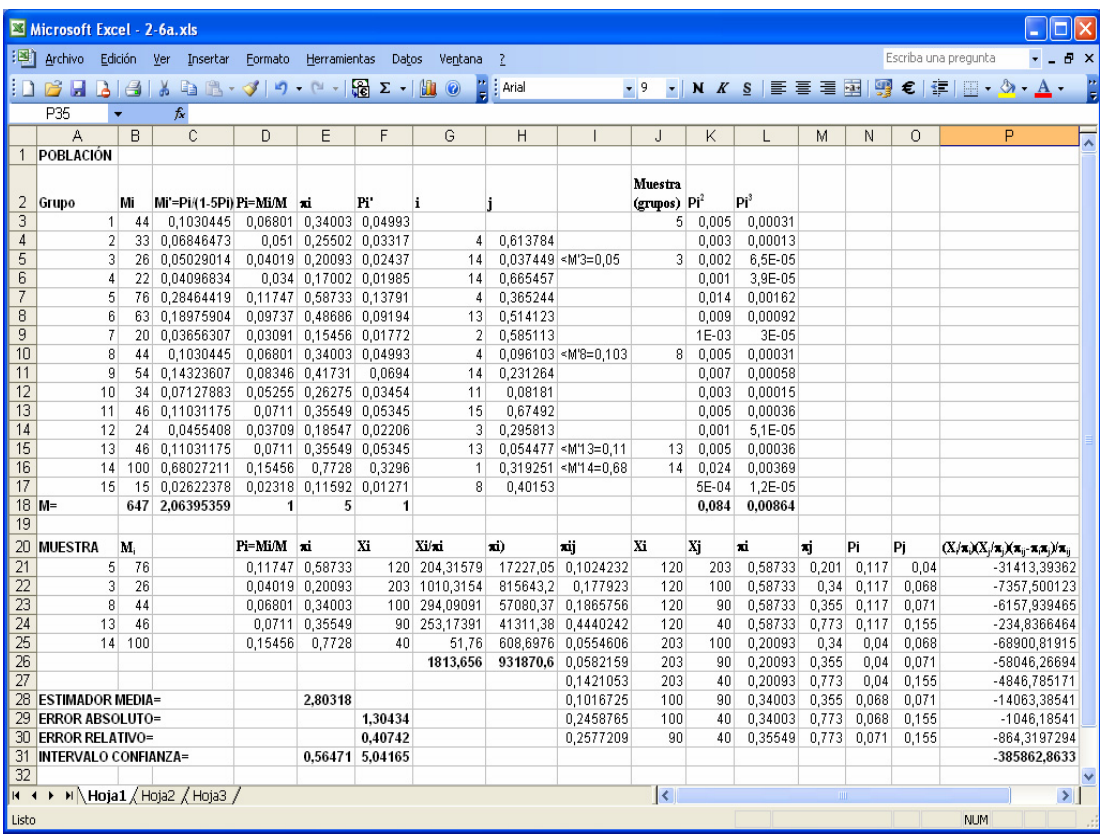
con lo que se tiene:

$$\hat{C}_v\left(\hat{\bar{X}}_{HT}\right)=\frac{\sqrt{\hat{V}\left(\hat{\bar{X}}_{HT}\right)}}{\hat{\bar{X}}_{HT}}=\frac{\sqrt{1{,}3043}}{2{,}8}=0{,}4074\rightarrow 41\%$$

Se observa que el error relativo de muestreo es del 41%. A continuación se realiza una estimación por intervalos al 95% de confianza.

$$\hat{\bar{X}}_{HT}\pm\lambda_\alpha\sqrt{\hat{V}(\hat{\bar{X}}_{HT})}=2{,}8\pm 1{,}96\sqrt{1{,}3043}=[0.56,\ 5.04]\rightarrow 95\%\ \textit{confianza}$$

A continuación se presentan los resultados automatizados con Excel.



Microsoft Excel - 2-6a.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	POBLACIÓN															
2	Grupo	Mi	Mi'=Pi/(1-5Pi)	Pi=Mi/M	xi	Pi'	i	j		Muestra (grupos)	Pi²	Pi³				
3	1	44	0,1030445	0,06801	0,34003	0,04993				5	0,005	0,00031				
4	2	33	0,06846473	0,051	0,25502	0,03317	4	0,613784			0,003	0,00013				
5	3	26	0,05029014	0,04019	0,20093	0,02437	14	0,037449	<M'3=0,05	3	0,002	6,5E-05				
6	4	22	0,04096834	0,034	0,17002	0,01985	14	0,665457			0,001	3,9E-05				
7	5	76	0,28464419	0,11747	0,58733	0,13791	4	0,365244			0,014	0,00162				
8	6	63	0,18975904	0,09737	0,48686	0,09194	13	0,514123			0,009	0,00092				
9	7	20	0,03656307	0,03091	0,15456	0,01772	2	0,585113			1E-03	3E-05				
10	8	44	0,1030445	0,06801	0,34003	0,04993	4	0,096103	<M'8=0,103	8	0,005	0,00031				
11	9	54	0,14323607	0,08346	0,41731	0,0694	14	0,231264			0,007	0,00058				
12	10	34	0,07127883	0,05255	0,26275	0,03454	11	0,08181			0,003	0,00015				
13	11	46	0,11031175	0,0711	0,35549	0,05345	15	0,67492			0,005	0,00036				
14	12	24	0,0455408	0,03709	0,18547	0,02206	3	0,295813			0,001	5,1E-05				
15	13	46	0,11031175	0,0711	0,35549	0,05345	13	0,054477	<M'13=0,11	13	0,005	0,00036				
16	14	100	0,68027211	0,15456	0,7728	0,3296	1	0,319251	<M'14=0,68	14	0,024	0,00369				
17	15	15	0,02622378	0,02318	0,11592	0,01271	8	0,40153			5E-04	1,2E-05				
18	M=	647	2,06395359	1	5	1					0,084	0,00864				
19																
20	MUESTRA	M_i		Pi=Mi/M	xi	Xi	Xi/xi	xi)	xij	Xi	Xj	xi	xj	Pi	Pj	$(X_i/\pi_i)(X_j/\pi_j)(\pi_{ij}-\pi_i\pi_j)/\pi_{ij}$
21	5	76		0,11747	0,58733	120	204,31579	17227,05	0,1024232	120	203	0,58733	0,201	0,117	0,04	-31413,39362
22	3	26		0,04019	0,20093	203	1010,3154	815643,2	0,177923	120	100	0,58733	0,34	0,117	0,068	-7357,500123
23	8	44		0,06801	0,34003	100	294,09091	57080,37	0,1865756	120	90	0,58733	0,355	0,117	0,071	-6157,939465
24	13	46		0,0711	0,35549	90	253,17391	41311,38	0,4440242	120	40	0,58733	0,773	0,117	0,155	-234,8366464
25	14	100		0,15456	0,7728	40	51,76	608,6976	0,0554606	203	100	0,20093	0,34	0,04	0,068	-68900,81915
26							1813,656	931870,6	0,0582159	203	90	0,20093	0,355	0,04	0,071	-58046,26694
27									0,1421053	203	40	0,20093	0,773	0,04	0,155	-4846,785171
28	ESTIMADOR MEDIA=				2,80318				0,1016725	100	90	0,34003	0,355	0,068	0,071	-14063,38541
29	ERROR ABSOLUTO=					1,30434			0,2458765	100	40	0,34003	0,773	0,068	0,155	-1046,18541
30	ERROR RELATIVO=					0,40742			0,2577209	90	40	0,35549	0,773	0,071	0,155	-864,3197294
31	INTERVALO CONFIANZA=				0,56471	5,04165										-385862,8633
32																

Hoja1 / Hoja2 / Hoja3

Listo NUM

2.7. Resolver el problema anterior suponiendo que se selecciona una muestra de tamaño 2 sin reposición mediante el método de Brewer.

Brewer propuso un método de selección para muestras de tamaño n tal que la primera unidad se extrae sin reposición con probabilidad proporcional al valor:

$$k_i = P_i \frac{(1-P_i)}{(1-2P_i)} \qquad P_i < ½$$

y el resto de las extracciones se realizan sin reposición y con probabilidades proporcionales a P_i.

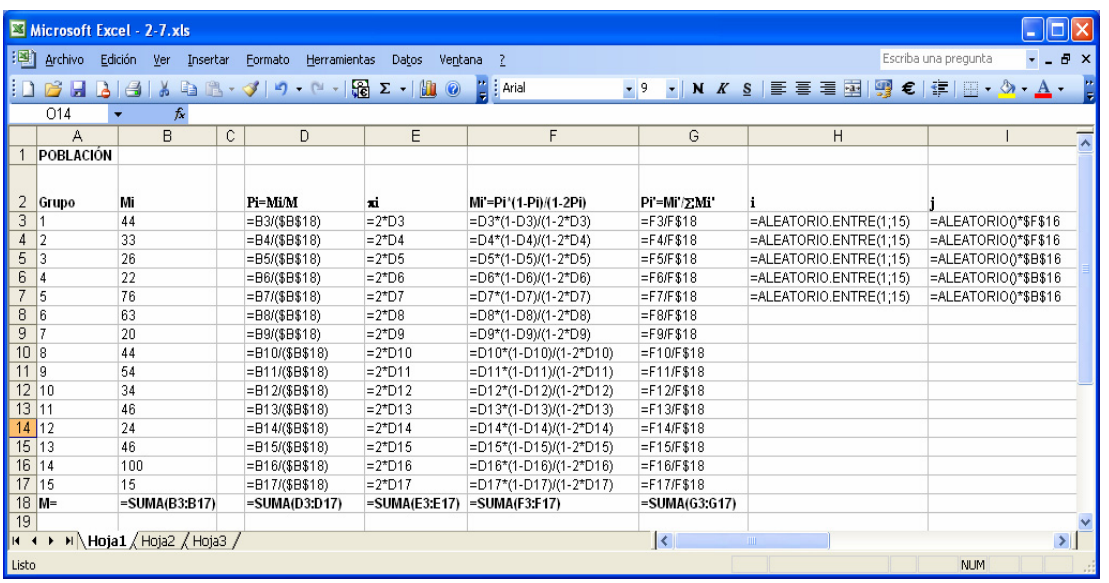
Para muestras de tamaño 2 las probabilidades π_i y π_{ij} son las siguientes:

$$\pi_i = 2P_i \qquad \pi_{ij} = \frac{2P_iP_j}{1+\sum_{i=1}^{N}\frac{P_i}{1-2P_i}} * \left[\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right]$$

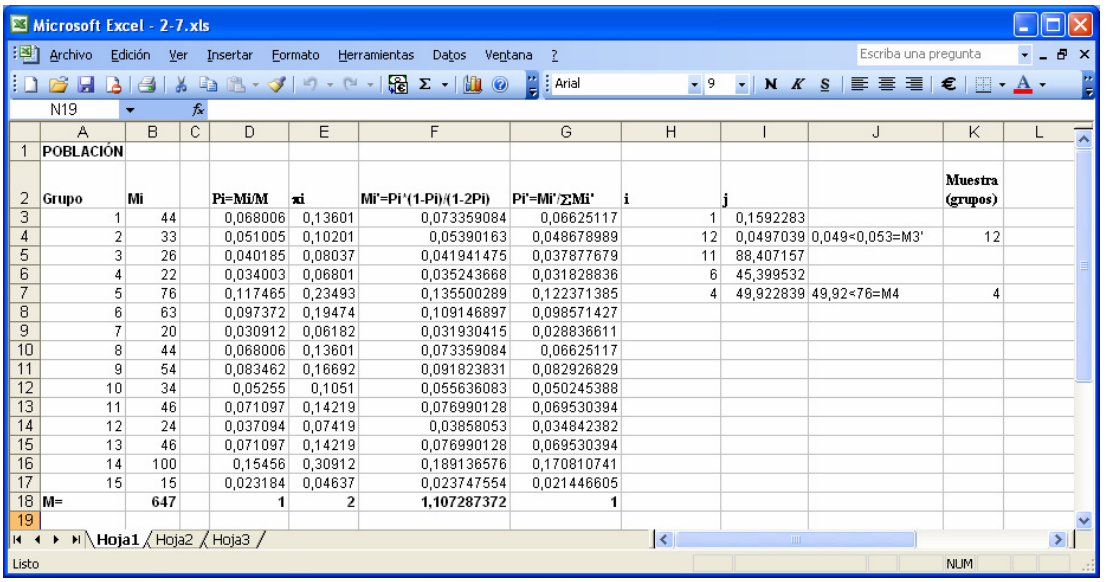
En nuestro problema, para realizar la primera extracción con probabilidades proporcionales a k_i aplicaremos el método de Lahiri seleccionando pares de números aleatorios (i,j) con $1 \le i \le 15$ y $0 \le j \le Máx(k_i)=0{,}68$ hasta que $j \le k_i$, con lo que resulta elegida la unidad u_{12} después de dos intentos.

Para realizar la segunda extracción con probabilidades proporcionales a M_i aplicaremos otra vez el método de Lahiri seleccionando pares de números aleatorios (i,j) con $1 \le i \le 15$ y $0 \le j \le Máx(M_i)=100$ hasta que $j \le M_i$, con lo que resulta elegida la unidad u_4 después de tres intentos.

Las tablas siguientes ilustran las fórmulas con Excel y los resultados obtenidos.



	A	B	C	D	E	F	G	H	I
1	POBLACIÓN								
2	Grupo	Mi		Pi=Mi/M	xi	Mi'=Pi'(1-Pi)/(1-2Pi)	Pi'=Mi'/ΣMi'	i	j
3	1	44		=B3/($B$18)	=2*D3	=D3*(1-D3)/(1-2*D3)	=F3/F$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$F$16
4	2	33		=B4/($B$18)	=2*D4	=D4*(1-D4)/(1-2*D4)	=F4/F$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$F$16
5	3	26		=B5/($B$18)	=2*D5	=D5*(1-D5)/(1-2*D5)	=F5/F$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$B$16
6	4	22		=B6/($B$18)	=2*D6	=D6*(1-D6)/(1-2*D6)	=F6/F$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$B$16
7	5	76		=B7/($B$18)	=2*D7	=D7*(1-D7)/(1-2*D7)	=F7/F$18	=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$B$16
8	6	63		=B8/($B$18)	=2*D8	=D8*(1-D8)/(1-2*D8)	=F8/F$18		
9	7	20		=B9/($B$18)	=2*D9	=D9*(1-D9)/(1-2*D9)	=F9/F$18		
10	8	44		=B10/($B$18)	=2*D10	=D10*(1-D10)/(1-2*D10)	=F10/F$18		
11	9	54		=B11/($B$18)	=2*D11	=D11*(1-D11)/(1-2*D11)	=F11/F$18		
12	10	34		=B12/($B$18)	=2*D12	=D12*(1-D12)/(1-2*D12)	=F12/F$18		
13	11	46		=B13/($B$18)	=2*D13	=D13*(1-D13)/(1-2*D13)	=F13/F$18		
14	12	24		=B14/($B$18)	=2*D14	=D14*(1-D14)/(1-2*D14)	=F14/F$18		
15	13	46		=B15/($B$18)	=2*D15	=D15*(1-D15)/(1-2*D15)	=F15/F$18		
16	14	100		=B16/($B$18)	=2*D16	=D16*(1-D16)/(1-2*D16)	=F16/F$18		
17	15	15		=B17/($B$18)	=2*D17	=D17*(1-D17)/(1-2*D17)	=F17/F$18		
18	M=	=SUMA(B3:B17)		=SUMA(D3:D17)	=SUMA(E3:E17)	=SUMA(F3:F17)	=SUMA(G3:G17)		
19									



	A	B	C	D	E	F	G	H	I	J	K	L
1	POBLACIÓN											
2	Grupo	Mi		Pi=Mi/M	πi	Mi'=Pi'(1-Pi)/(1-2Pi)	Pi'=Mi'/ΣMi'	i	j		Muestra (grupos)	
3	1	44		0,068006	0,13601	0,073359084	0,06625117	1	0,1592283			
4	2	33		0,051005	0,10201	0,05390163	0,048678989	12	0,0497039	0,049<0,053=M3'	12	
5	3	26		0,040185	0,08037	0,041941475	0,037877679	11	88,407157			
6	4	22		0,034003	0,06801	0,035243668	0,031828836	6	45,399532			
7	5	76		0,117465	0,23493	0,135500289	0,122371385	4	49,922839	49,92<76=M4	4	
8	6	63		0,097372	0,19474	0,109146897	0,098571427					
9	7	20		0,030912	0,06182	0,031930415	0,028836611					
10	8	44		0,068006	0,13601	0,073359084	0,06625117					
11	9	54		0,083462	0,16692	0,091823831	0,082926829					
12	10	34		0,05255	0,1051	0,055636083	0,050245388					
13	11	46		0,071097	0,14219	0,076990128	0,069530394					
14	12	24		0,037094	0,07419	0,03858053	0,034842382					
15	13	46		0,071097	0,14219	0,076990128	0,069530394					
16	14	100		0,15456	0,30912	0,189136576	0,170810741					
17	15	15		0,023184	0,04637	0,023747554	0,021446605					
18	M=	647		1	2	1,107287372	1					
19												

La muestra estará formada por los grupos {12, 4} cuyos tamaños son los siguientes:

$Grupos\ (Muestra)$	12	4
$Tamaños\ (M_i)$	24	22

y el número total de horas semanales empleadas por los estudiantes de esos grupos para estudiar la materia Introducción a la Estadística es el siguiente:

$Grupos\ (Muestra)$	12	4
$Horas\ (X_i)$	120	203

A continuación se realiza la estimación del promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística utilizando el estimador de Horvitz y Thompson (ya que el método de selección de la muestra es sin reposición). En la tabla siguiente se presentan todos los cálculos necesarios para realizar la estimación ($N = 15$, $n = 2$).

MUESTRA	M_i		Pi=Mi/M	πi	Xi	Xi/πi	$(Xi/\pi i)^2(1-\pi i)$
12	24		0,03709	0,074	120	1617,5	2422206,3
4	22		0,034	0,068	203	2985,0227	8304401,1
						4602,5227	**10726607**

$$\hat{\bar{X}}_{HT} = \frac{1}{M}\hat{X}_{HT} = \frac{1}{M}\sum_{i=1}^{25}\frac{X_i}{\pi_i} = \frac{1}{647}\left(\frac{120}{0{,}074} + \frac{203}{0{,}068}\right) = \frac{4602{,}5227}{647} = 4{,}614$$

Por lo tanto, se estima que el promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística es 4,6 horas. A continuación hallamos el error de esta estimación a través de la estimación de la varianza. Ahora se presenta la tabla con todos los cálculos necesarios para realizar la estimación ($N = 15$, $n = 2$).

$(Xi/\pi i)^2(1-\pi i)$	πij	Xi	Xj	πi	πj	Pi	Pj	$(X_i/\pi_i)(X_j/\pi_j)(\pi_{ij}-\pi_i\pi_j)/\pi_{ij}$
2422206,3	0,00297	120	203	0,07419	0,068	0,037	0,034	-3383694,955
8304401,1								**-6767389,909**
10726607								

$$\hat{V}\left(\hat{\bar{X}}_{HT}\right)=\frac{1}{M^2}\hat{V}\left(\hat{X}_{HT}\right)=\frac{1}{M^2}\left[\sum_{i=1}^{5}\frac{X_i^2}{\pi_i^2}(1-\pi_i)+2\sum_{i=1}^{5}\sum_{j>i}^{5}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\right]=$$

$$\frac{1}{M^2}\left[\frac{X_1^2}{\pi_1^2}(1-\pi_1)+\frac{X_2^2}{\pi_2^2}(1-\pi_2)+2\left(\frac{X_1}{\pi_1}\frac{X_2}{\pi_2}\left(\frac{\pi_{12}-\pi_1\pi_2}{\pi_{12}}\right)\right)\right]=9{,}458$$

con lo que se tiene:

$$\hat{C}_v\left(\hat{\bar{X}}_{HT}\right)=\frac{\sqrt{\hat{V}\left(\hat{\bar{X}}_{HT}\right)}}{\hat{\bar{X}}_{HT}}=\frac{\sqrt{9{,}458}}{4{,}614}=0{,}66\rightarrow 66\%$$

Se observa que el error relativo de muestreo es del 66%. A continuación se realiza una estimación por intervalos al 95% de confianza.

$$\hat{\bar{X}}_{HT}\pm\lambda_\alpha\sqrt{\hat{V}(\hat{\bar{X}}_{HT})}=4{,}614\pm 1{,}96\sqrt{9{,}458}=[-1.414,\ 10.6414]\rightarrow 95\%\ \textit{confianza}$$

A continuación se presentan los resultados automatizados con Excel.

Microsoft Excel - 2-7a.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	POBLACIÓN															
2	Grupo	Mi		Pi=Mi/M	xi	ki=Pi'(1-Pi)/(1-2Pi)	Pi'=ki'/Σki'	i	j		Muestra (grupos)	Pi(1-2Pi)				
3	1	44		0,06801	0,136	0,073359084	0,0662512	1	0,15923			0,05876				
4	2	33		0,051	0,102	0,05390163	0,048679	12	0,0497	0,049<0,053=M3'	12	0,0458				
5	3	26		0,04019	0,08	0,041941475	0,0378777	11	88,4072			0,03696				
6	4	22		0,034	0,068	0,035243668	0,0318288	6	45,3995			0,03169				
7	5	76		0,11747	0,235	0,135500289	0,1223714	4	49,9228	49,92<76=M4	4	0,08987				
8	6	63		0,09737	0,195	0,109146897	0,0985714					0,07841				
9	7	20		0,03091	0,062	0,031930415	0,0288366					0,029				
10	8	44		0,06801	0,136	0,073359084	0,0662512					0,05876				
11	9	54		0,08346	0,167	0,091823831	0,0829268					0,06953				
12	10	34		0,05255	0,105	0,055636083	0,0502454					0,04703				
13	11	46		0,0711	0,142	0,076990128	0,0695304					0,06099				
14	12	24		0,03709	0,074	0,03858053	0,0348424					0,03434				
15	13	46		0,0711	0,142	0,076990128	0,0695304					0,06099				
16	14	100		0,15456	0,309	0,189136576	0,1708107					0,10678				
17	15	15		0,02318	0,046	0,023747554	0,0214466					0,02211				
18	M=	647		1	2	1,10728737	1					0,83101				
19																
20	MUESTRA	M_i		Pi=Mi/M	xi	Xi	Xi/xi	(Xi/xi)²(1-xi)	xij	Xi	Xj	xi	xj	Pi	Pj	$(X_i/\pi_i)(X_j/\pi_j)(\pi_{ij}-\pi_i\pi_j)/\pi_{ij}$
21	12	24		0,03709	0,074	120	1617,5	2422206,3	0,00297	120	203	0,07419	0,068	0,037	0,034	-3383694,955
22	4	22		0,034	0,068	203	2985,0227	8304401,1								-6767389,909
23							4602,5227	10726607								
24																
25	ESTIMADOR MEDIA=				4,614											
26	ERROR ABSOLUTO=					9,45803222										
27	ERROR RELATIVO=					0,666587297										
28	INTERVALO CONFIANZA=				-1,414	10,64140349										
29																
30																

Hoja1 / Hoja2 / Hoja3

Listo NUM

2.8.

Resolver el problema anterior suponiendo que se selecciona una muestra de tamaño 2 sin reposición utilizando el estimador de Murthy.

Murthy mejoró un método anterior de Des Raj extrayendo unidades sucesivas para la muestra con probabilidades P_i, $P_j(1-\underline{P_i})$, $P_k(1-P_i-P_j)$ y así sucesivamente. Propuso el estimador del total:

$$\hat{X}_M = \frac{\sum_{i=1}^{n} P(S/i)X_i}{P(S)}, \quad \hat{V}\left(\hat{X}_M\right) = \frac{1}{P(S)^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\left[P(S)P(S/i,j) - P(S/i)P(S/j)\right]P_iP_j\left(\frac{X_i}{P_i} - \frac{X_j}{P_j}\right)^2$$

P(S) = Probabilidad incondicional de obtener la muestra *S*.
P(S/i) = Probabilidad de obtener la muestra *S* condicionado a que se sacó la unidad *i* la primera
P(S/i,j) = Probabilidad de *S* condicionado a que se sacaron las unidades *i* y *j* las dos primeras.

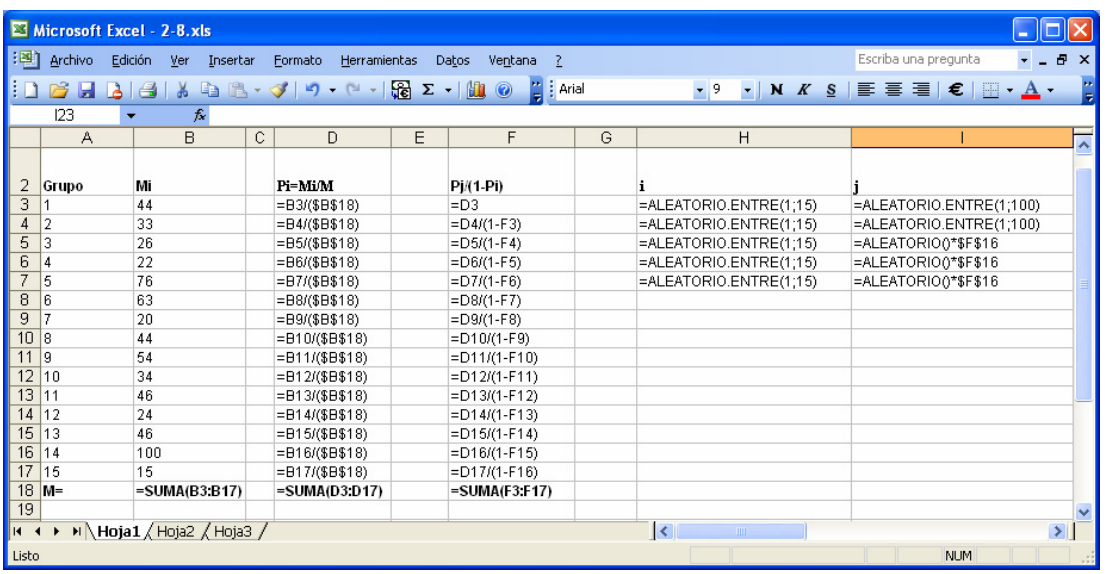
Para n=2 se tiene que $P(S/i) = P_j/(1-P_i)$ y $P(S) = \pi_{ij} = P_iP_j(2-P_i-P_j)/(1-P_i)(1-P_j)$ y además:

$$\pi_i = P_i\left[1+\sum_{j\neq i}\frac{P_j}{1-P_j}\right] \quad \hat{X}_M = \frac{1}{2-P_i-P_j}\left[(1-P_j)\frac{X_i}{P_i} + (1-P_i)\frac{X_j}{P_j}\right], \quad \hat{V}(\hat{X}_M) = \frac{(1-P_i)(1-P_j)(1-P_i-P_j)}{(2-P_i-P_j)^2}\left(\frac{X_i}{P_i} - \frac{X_j}{P_j}\right)^2$$

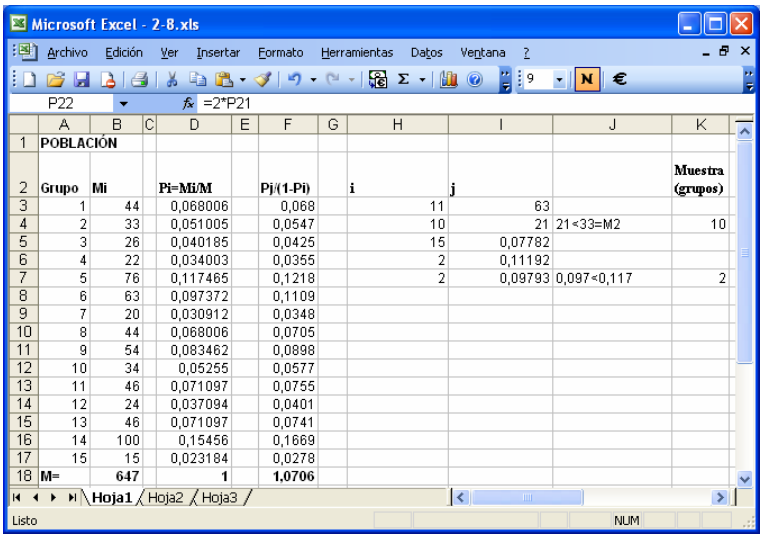
En nuestro problema, para realizar la primera extracción con probabilidad P_i proporcional a su tamaño M_i, aplicamos el método de Lahiri seleccionando pares de números aleatorios (i,j) con $1 \le i \le 15$ y $0 \le j \le Máx(M_i)$=100 hasta que $j \le M_i$, con lo que resulta elegida la unidad u_{10} después de dos intentos.

Para realizar la segunda extracción con probabilidad $P_j/(1-P_i)$ aplicamos otra vez el método de Lahiri seleccionando pares de números aleatorios (i,j) con $1 \le i \le 15$ y $0 \le j \le Máx(P_j/(1-P_i)) = 0{,}15456$ hasta que $j \le P_j/(1-P_i)$, con lo que resulta elegida la unidad u_2 después de tres intentos.

Las tablas siguientes ilustran las fórmulas con Excel y los resultados obtenidos.



	A	B	C	D	E	F	G	H	I
2	**Grupo**	**Mi**		**Pi=Mi/M**		**Pj/(1-Pi)**		**i**	**j**
3	1	44		=B3/($B$18)		=D3		=ALEATORIO.ENTRE(1;15)	=ALEATORIO.ENTRE(1;100)
4	2	33		=B4/($B$18)		=D4/(1-F3)		=ALEATORIO.ENTRE(1;15)	=ALEATORIO.ENTRE(1;100)
5	3	26		=B5/($B$18)		=D5/(1-F4)		=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$F$16
6	4	22		=B6/($B$18)		=D6/(1-F5)		=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$F$16
7	5	76		=B7/($B$18)		=D7/(1-F6)		=ALEATORIO.ENTRE(1;15)	=ALEATORIO()*$F$16
8	6	63		=B8/($B$18)		=D8/(1-F7)			
9	7	20		=B9/($B$18)		=D9/(1-F8)			
10	8	44		=B10/($B$18)		=D10/(1-F9)			
11	9	54		=B11/($B$18)		=D11/(1-F10)			
12	10	34		=B12/($B$18)		=D12/(1-F11)			
13	11	46		=B13/($B$18)		=D13/(1-F12)			
14	12	24		=B14/($B$18)		=D14/(1-F13)			
15	13	46		=B15/($B$18)		=D15/(1-F14)			
16	14	100		=B16/($B$18)		=D16/(1-F15)			
17	15	15		=B17/($B$18)		=D17/(1-F16)			
18	**M=**	**=SUMA(B3:B17)**		**=SUMA(D3:D17)**		**=SUMA(F3:F17)**			
19									



	A	B	C	D	E	F	G	H	I	J	K
1	POBLACIÓN										
2	Grupo	Mi		Pi=Mi/M		Pj/(1-Pi)		i	j		Muestra (grupos)
3	1	44		0,068006		0,068		11	63		
4	2	33		0,051005		0,0547		10	21	21<33=M2	10
5	3	26		0,040185		0,0425		15	0,07782		
6	4	22		0,034003		0,0355		2	0,11192		
7	5	76		0,117465		0,1218		2	0,09793	0,097<0,117	2
8	6	63		0,097372		0,1109					
9	7	20		0,030912		0,0348					
10	8	44		0,068006		0,0705					
11	9	54		0,083462		0,0898					
12	10	34		0,05255		0,0577					
13	11	46		0,071097		0,0755					
14	12	24		0,037094		0,0401					
15	13	46		0,071097		0,0741					
16	14	100		0,15456		0,1669					
17	15	15		0,023184		0,0278					
18	M=	647		1		1,0706					

La muestra estará formada por los grupos $\{10, 2\}$ cuyos tamaños son los siguientes:

$Grupos\ (Muestra)$	10	2
$Tamaños\ (M_i)$	34	33

y el número total de horas semanales empleadas por los estudiantes de esos grupos para estudiar la materia Introducción a la Estadística es el siguiente:

$Grupos\ (Muestra)$	10	2
$Horas\ (X_i)$	120	203

A continuación se realiza la estimación del promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística utilizando el estimador de Murthy. En la siguiente tabla se presentan todos los cálculos necesarios para realizar la estimación ($N = 15$, $n = 2$).

MUESTRA	M_i	Pi=Mi/M	πi	Xi	Xi/Pi	ESTIMADOR DEL TOTAL
10	34	0,05255	0,055375	120	2283,5294	**3131,088537**
2	33	0,051005	0,053834	203	3980,0303	

$$\hat{\bar{X}}_{HT} = \frac{1}{M}\hat{X}_{HT} = \frac{1}{M}\frac{1}{2 - P_i - P_j}\left[(1 - P_j)\frac{X_i}{P_i} + (1 - P_i)\frac{X_j}{P_j}\right] = \frac{1}{647}3131{,}088 = 4{,}839$$

Por lo tanto, se estima que el promedio de horas semanales que dedican los estudiantes a la materia de Introducción a la Estadística es 4,8 horas. A continuación hallamos el error de esta estimación a través de la estimación de la varianza.

$$\hat{V}\left(\hat{\bar{X}}_{HT}\right) = \frac{1}{M^2}\hat{V}\left(\hat{X}_{HT}\right) = \frac{1}{M^2}\left[\frac{(1 - P_i)(1 - P_j)(1 - P_i - P_j)}{(2 - P_i - P_j)^2}\left(\frac{X_i}{P_i} - \frac{X_j}{P_j}\right)^2\right] = \frac{1}{647^2}1223240{,}6 = 2{,}92$$

con lo que se tiene:

$$\hat{C}_v\left(\hat{\bar{X}}_{HT}\right)=\frac{\sqrt{\hat{V}\left(\hat{\bar{X}}_{HT}\right)}}{\hat{\bar{X}}_{HT}}=\frac{\sqrt{2{,}92}}{4{,}839}=0{,}3532\rightarrow 35\%$$

Se observa que el error relativo de muestreo es del 35%. A continuación se realiza una estimación por intervalos al 95% de confianza.

$$\hat{\bar{X}}_{HT}\pm\lambda_\alpha\sqrt{\hat{V}(\hat{\bar{X}}_{HT})}=4{,}839\pm 1{,}96\sqrt{2{,}92}=[-1.48,\ 8.18]\rightarrow 95\%\ \textit{confianza}$$

A continuación se presentan los resultados y fórmulas con Excel.

Microsoft Excel - 2-8.xls

	A	B	C	D	E	F	G	H	I	J	K
1	POBLACIÓN										
2	Grupo	Mi		Pi=Mi/M		Pj/(1-Pi)		i	j		Muestra (grupos)
3	1	44		0,068006		0,06801		11	63		
4	2	33		0,051005		0,05473		10	21	21<33=M2	10
5	3	26		0,040185		0,04251		15	0,07782		
6	4	22		0,034003		0,03551		2	0,11192		
7	5	76		0,117465		0,12179		2	0,09793	0,097<0,117	2
8	6	63		0,097372		0,11088					
9	7	20		0,030912		0,03477					
10	8	44		0,068006		0,07046					
11	9	54		0,083462		0,08979					
12	10	34		0,05255		0,05773					
13	11	46		0,071097		0,07545					
14	12	24		0,037094		0,04012					
15	13	46		0,071097		0,07407					
16	14	100		0,15456		0,16692					
17	15	15		0,023184		0,02783					
18	M=	647		1		1,07057					
19											
20	MUESTRA	M_i		Pi=Mi/M	$\bar{x}_i$	Xi	Xi/Pi	ESTIMADOR DEL TOTAL	$\bar{x}_{ij}$	VARIANZA DEL TOTAL	
21	10	34		0,05255	0,055375	120	2283,5294	3131,088537	0,0056533	1223240,635	
22	2	33		0,051005	0,053834	203	3980,0303				
23											
24											
25	ESTIMADOR MEDIA=				4,839395						
26	ERROR ABSOLUTO=					2,92216					
27	ERROR RELATIVO=					0,35323					
28	INTERVALO CONFIANZA=				1,488909	8,18988					
29											

Hoja1 / Hoja2 / Hoja3

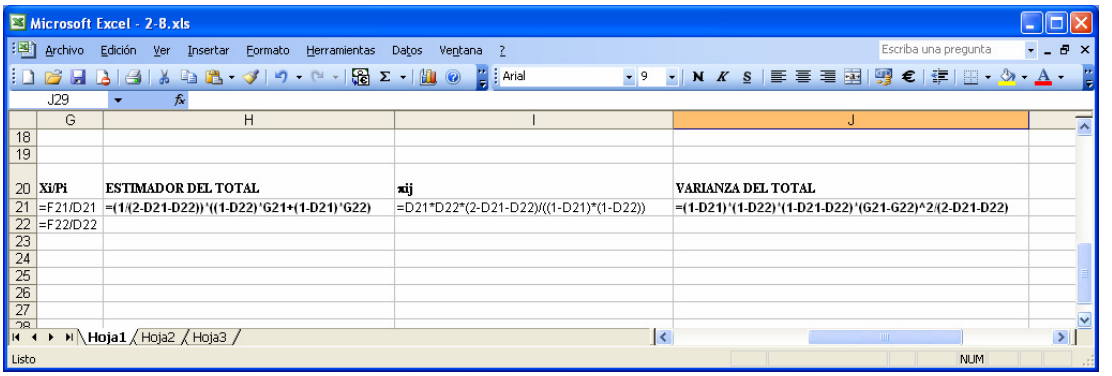
Listo NUM

Microsoft Excel - 2-8.xls

	A	B	C	D	E	F	G	H
20	MUESTRA	M_i		Pi=Mi/M	$\bar{x}_i$	Xi	Xi/Pi	ESTIMADOR DEL TOTAL
21	10	34		=B21/$B$18	=D21*(1+D22/(1-D22))	120	=F21/D21	=(1/(2-D21-D22))*((1-D22)*G21+(1-D21)*G22)
22	2	33		=B22/$B$18	=D22*(1+D21/(1-D21))	203	=F22/D22	
23								
24								
25	ESTIMADOR MEDIA=				=H21/B18			
26	ERROR ABSOLUTO=					=J21/B18^2		
27	ERROR RELATIVO=					=RAIZ(F26)/E25		
28	INTERVALO CONFIANZA=				=$E$25-1,96*RAIZ($F$26)	=$E$25+1,96*RAIZ($F$26)		
29								

Hoja1 / Hoja2 / Hoja3

Listo NUM



2.9. En una región montañosa de 25000 hectáreas se trata de estudiar la superficie dedicada a la plantación de pinos. La región de divide en 100 zonas disjuntas lo más similares entre sí, de tal forma que cada zona contiene plantas de todas las clases que crecen en la región. Se extrae una muestra de 10 zonas con reemplazamiento y con probabilidades proporcionales a sus superficies. Las proporciones de superficie total dedicadas a la plantación de pinos en cada una de las zonas de la muestra son:

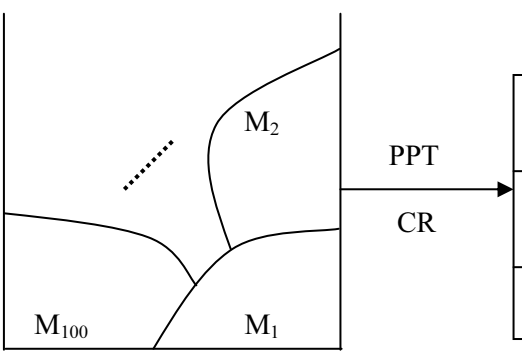
0.05, 0.25, 0.10, 0.30, 0.15, 0.25, 0.35, 0.25, 0.10 y 0.20

Se pide:

1) Un estimador insesgado de la superficie total de la región dedicada a la plantación de pinos, su error relativo y un intervalo de confianza al nivel $\alpha = 0{,}05$.

2) Contestar a las mismas preguntas del apartado anterior suponiendo que la selección es sin reposición mediante el método de Ikeda. En este caso considerar la muestra con sólo tres zonas de igual superficie (250 hectáreas) para las que las proporciones de superficie total dedicadas a la plantación de pinos en cada una de ellas son 0.25, 0.35 y 0.40, respectivamente. Se supone en este caso que las 100 zonas de la población son de igual superficie.

Considerando muestreo con reposición (CR) y selección con probabilidades proporcionales a los tamaños (PPT), el esquema del problema es el siguiente:



PPT

CR

X_1/M_1 0,05	X_2/M_2 0,25	X_3/M_3 0,10	
X_4/M_4 0,30	X_5/M_5 0,15	X_6/M_6 0,25	X_7/M_7 0,35
X_8/M_8 0,25	X_9/M_9 0,10	X_{10}/M_{10} 0,20	

N = 100 M = 25000 n = 20

Sea M_i = Superficie de la zona i-ésima
Sea X_i = Superficie dedicada a la plantación de pinos

$$\hat{X}_{HH} = \sum_{i=1}^{n} \frac{X_i}{nP_i} = \sum_{i=1}^{n} \frac{X_i}{n\dfrac{M_i}{M}} = \frac{M}{n}\sum_{i=1}^{n} \frac{X_i}{M_i} = \frac{25000}{10}(0{,}05 + 0{,}25 + \cdots + 0{,}20) = 5000$$

$$\hat{V}(\hat{X}_{HH}) = \frac{\sum_{i=1}^{n}\left(\dfrac{X_i}{P_i} - \hat{X}_{HH}\right)^2}{n(n-1)} = \frac{\sum_{i=1}^{n}\left(\dfrac{X_i}{M_i/M} - \hat{X}_{HH}\right)^2}{n(n-1)} = \frac{\sum_{i=1}^{n}\left(M\dfrac{X_i}{M_i} - \hat{X}_{HH}\right)^2}{n(n-1)} =$$

$$\frac{(25000\cdot 0{,}05 - 5000)^2 + (25000\cdot 0{,}25 - 5000)^2 + \cdots + (25000\cdot 0{,}20 - 5000)^2}{10(10-1)} = 590278$$

$$\hat{C}v(\hat{X}) = \frac{\sqrt{\hat{V}(\hat{X})}}{\hat{X}} = \frac{\sqrt{590278}}{5000} = 0{,}15 \quad (15\%)$$

La estimación por intervalos suponiendo normalidad en la población es:

$$\hat{X} \pm \lambda_\alpha \hat{\sigma}(\hat{X}) = 5000 \pm 2\sqrt{590278} = [3464,\ 6536]$$

La estimación por intervalos sin normalidad en la población es:

$$\hat{X} \pm \frac{\hat{\sigma}(\hat{X})}{\sqrt{\alpha}} = 5000 \pm \sqrt{\frac{590278}{0{,}05}} = [1564,\ 8346]$$

Para resolver el segundo apartado del problema consideramos la muestra con sólo tres zonas de igual superficie ($M_1 = M_2 = M_3 = 250$) para las que las proporciones de superficie total dedicadas a la plantación de pinos en cada una de ellas son de 0,25, 0,35 y 0,40, respectivamente. Como los P_i son proporcionales a las superficies de las zonas se tiene:

$$\left.\begin{aligned} \frac{X_1}{M_1} = \frac{X_1}{250} = 0{,}25 \Rightarrow X_1 = 62{,}5 \\ \frac{X_2}{M_2} = \frac{X_2}{250} = 0{,}35 \Rightarrow X_2 = 87{,}5 \\ \frac{X_3}{M_3} = \frac{X_3}{250} = 0{,}40 \Rightarrow X_3 = 100 \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} P_i &= \frac{M_i}{M} = \frac{250}{25000} = 0{,}01 \quad (i = 1,2,3 \quad j = 1,2,3) \\ \pi_i &= P_i + \frac{n-1}{N-1}(1 - P_i) = 0{,}01 + \frac{2}{99}\cdot 0{,}9 = 0{,}028 \\ \pi_{ij} &= \frac{(n-1)}{(N-1)}\left[(N-n)\cdot\frac{P_i + P_j}{N-2} + \frac{n-2}{N-2}\right] = \\ &= \frac{(3-1)}{(100-1)}\left[(100-3)\cdot\frac{0{,}02}{100-2} + \frac{3-2}{100-2}\right] = 0{,}006 \end{aligned}\right.$$

$$\text{Sin reposición} \Rightarrow \hat{X}_{HT} = \sum_{i=1}^{n} \frac{X_i}{\pi_i} = \frac{1}{0{,}028}(62{,}5 + 87{,}5 + 100) = 8928{,}6$$

$$\hat{V}(\hat{X}_{HT}) = \sum_{i=1}^{n}\left(\frac{X_i}{\pi_i}\right)^2(1-\pi_i) + 2\sum_{i<j}^{n}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right) = 49429600$$

$$\hat{Cv}(\hat{X}) = \frac{\sqrt{49429600}}{8928{,}6} = 0{,}78 \qquad \hat{X} \pm \lambda_\alpha \hat{\sigma}(\hat{X}) = \left[-5122.6,\ 22989.8\right]$$

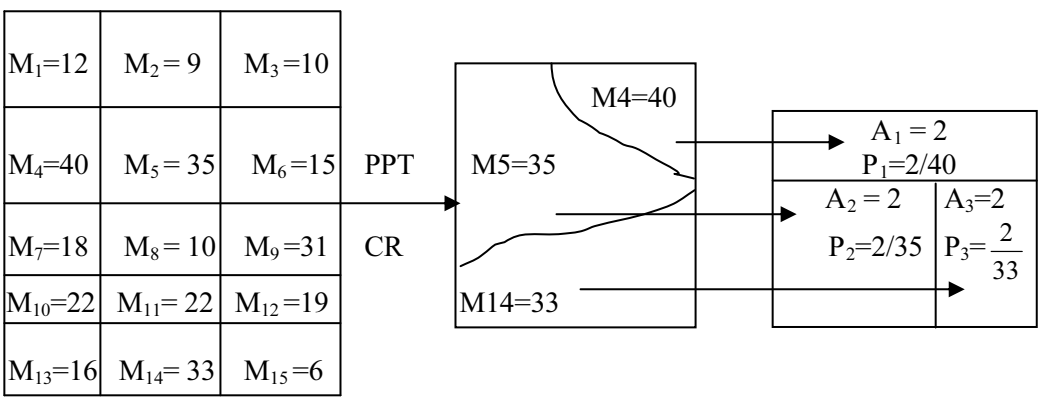
2.10. Una gran empresa tiene sus inventarios de equipo listados separadamente en 15 departamentos. Se selecciona una muestra de tres departamentos con reposición y probabilidades proporcionales al número de artículos de equipo en cada departamento. La tabla siguiente presenta el número de artículos de equipo NA en cada departamento D.

D	*NA*	*D*	*NA*	*D*	*NA*	*D*	*NA*	*D*	*NA*
1	12	4	40	7	18	10	22	13	16
2	9	5	35	8	10	11	22	14	33
3	27	6	15	9	31	12	19	15	6

a) Suponiendo que los tres departamentos seleccionados (que serán los de mayor probabilidad) tienen cada uno 2 artículos impropiamente identificados, estimar el número total de artículos impropiamente identificados en la empresa y su error relativo de muestreo.

b) Estimar por intervalos al 95% la media de artículos propiamente identificados, sabiendo que los tres departamentos seleccionados tienen respectivamente 4, 5 y 6 artículos impropiamente identificados. ¿Qué estimador es mejor?

El esquema del problema es el siguiente:



N = 15, M=315 n = 3

Como se selecciona la muestra de tres departamentos con probabilidades PPT proporcionales al número de artículos de equipo en cada departamento, los tres departamentos seleccionados para la muestra serán el 4, el 5 y el 14, ya que son los que van a tener mayor probabilidad de selección (por tener el mayor número de artículos). Al ser la selección con probabilidades proporcionales a los tamaños, se tiene que $P_i = M_i / M$, con lo que:

$$P_1 = \frac{40}{315},\ P_2 = \frac{35}{315} \text{ y } P_3 = \frac{33}{315}$$

Como el muestreo es con reposición, el estimador insesgado del total de la clase de los artículos impropiamene clasificados vendrá dado por la fórmula de Hansen y Hurwitz.

$$\hat{A}_{HH} = \sum_i^n \frac{A_i}{nP_i} = \sum_i^n \frac{M_i P_i}{n M_i/M} = \frac{1}{n}\sum_i^n \frac{M_i P_i}{M_i/M} = \frac{M}{n}\sum_i^n P_i = \frac{315}{3}\left(\frac{2}{40}+\frac{2}{35}+\frac{2}{33}\right) \cong 18$$

$\hat{P}_i$ = proporción muestral en el conglomerado *i*-ésimo.

Como estamos en muestreo con reposición y probabilidades desiguales proporcionales a los tamaños, utilizamos para estimar la varianza la siguiente expresión:

$$\hat{V}\left(\hat{A}\right) = \frac{\sum_i^n \left(\frac{A_i}{P_i} - \hat{A}\right)^2}{n(n-1)} = \frac{\sum_i^n \left(\frac{M_i P_i}{P_i} - M\hat{P}\right)^2}{n(n-1)} = \frac{M^2 \sum_i^n \left(P_i - \hat{P}\right)^2}{n(n-1)} =$$

$$\frac{315^2}{3\cdot 2}\left[\left(\frac{2}{40}-\frac{18}{315}\right)^2 + \left(\frac{2}{35}-\frac{18}{315}\right)^2 + \left(\frac{2}{33}-\frac{18}{315}\right)^2\right] = 1{,}04209$$

Para estimar la proporción de artículos propiamente identificados observamos que los tres departamentos seleccionados para la muestra (el 4, el 5 y el 14) tienen 36, 30 y 27 artículos propiamente identificados respectivamente. El estimador será el siguiente:

$$\hat{P}_{HH} = \frac{1}{M}\hat{A}_{HH} = \frac{1}{M}\sum_i^n \frac{A_i}{nP_i} = \frac{1}{M}\left(\frac{M}{n}\sum_i^n P_i\right) = \frac{1}{n}\sum_i^n P_i = \frac{1}{3}\left(\frac{36}{40}+\frac{30}{35}+\frac{27}{33}\right) = 0{,}858$$

$$\hat{V}\left(\hat{P}\right) = \frac{1}{M^2}\hat{V}\left(\hat{A}\right) = \frac{\sum_i^n \left(P_i - \hat{P}\right)^2}{n(n-1)} = \frac{1}{3\cdot 2}\left[\left(\frac{36}{40}-0{,}858\right)^2 + \left(\frac{30}{35}-0{,}858\right)^2 + \left(\frac{27}{33}-0{,}858\right)^2\right] = 0{,}000558$$

El intervalo de confianza al 95%, suponiendo normalidad, será:

$$\hat{\hat{P}} \pm \lambda_\alpha \sqrt{\hat{V}(\hat{\hat{P}})} = 0{,}858 \pm 1{,}96\sqrt{0{,}000558} = [0.8117,\ 0.9043]$$

2.11. Un gran banco que tiene 1000 sucursales con cuarenta microordenadores en cada una, emprende un proceso de auditoría informática. Para ello se extrae una muestra sin reposición y probabilidades iguales de 20 sucursales, resultando que en nueve de ellas no hay microordenadores con defectos, en ocho hay un ordenador defectuoso y en tres hay dos ordenadores defectuosos. Se pide:

1) Estimar el número total de microordenadores defectuosos en el banco y sus errores absoluto y relativo de muestreo. Realizar la estimación por intervalos al 99% ($F^{-1}(0{,}995) = 2{,}57$).

2) Resolver el problema con reposición y comparar los resultados con los del apartado primero.

Tenemos como datos $N = 1000$, $M = 40000$ y $n = 20$. Como el muestreo es sin reposición, el total de microordenadores defectuosos puede estimarse mediante el estimador de Horvitz y Thompson. Además, al ser el muestreo con probabilidades iguales tenemos que $\pi_i = n/N = 20/1000 = 0{,}02$ y $\pi_{ij} = 20(20\text{-}1)/[1000(1000\text{-}1)] = 0{,}00038$. Se tiene:

$$\hat{A}_{HT} = \sum_{i=1}^{25} \frac{A_i}{\pi_i} = \frac{9 \cdot 0 + 8 \cdot 1 + 3 \cdot 2}{0{,}02} = 700$$

La varianza se estima de la siguiente forma:

$$\hat{V}\left(\hat{A}_{HT}\right) = \sum_{i=1}^{20} \frac{A_i^2}{\pi_i^2}(1-\pi_i) + 2\sum_{i=1}^{20}\sum_{j>i}^{20} \frac{A_i}{\pi_i}\frac{A_j}{\pi_j}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right) = \frac{1-0{,}02}{0{,}02^2}\sum_{i=1}^{20} A_i^2 + \frac{2(0{,}00038-0{,}02^2)}{0{,}02^2 \cdot 0{,}00038}\sum_{i=1}^{20}\sum_{j>i}^{20} A_i A_j$$

$$2450(9 \cdot 0^2 + 8 \cdot 1^2 + 3 \cdot 2^2) - 263{,}15\left[\binom{9}{2}(0 \cdot 0) + 9 \cdot 8(0 \cdot 1) + 9 \cdot 3(0 \cdot 2) + \binom{8}{2}(1 \cdot 1) + 8 \cdot 3(1 \cdot 2) + \binom{3}{2}(2 \cdot 2)\right]$$

$$= 25842{,}1$$

Ahora calculamos el error relativo.

$$\hat{C}v(\hat{A}) = \frac{\sqrt{\hat{V}(\hat{A})}}{\hat{A}} = \frac{\sqrt{25842{,}1}}{700} = 0{,}2296 \quad (22{,}96\%)$$

La estimación por intervalos suponiendo normalidad en la población es:

$$\hat{A} \pm \lambda_\alpha \hat{\sigma}(\hat{A}) = 700 \pm 2.57\sqrt{25842{,}1} = \left[286.86,\ 1113.14\right]$$

La estimación por intervalos sin normalidad en la población es:

$$\hat{A} \pm \frac{\hat{\sigma}(\hat{A})}{\sqrt{\alpha}} = 700 \pm \sqrt{\frac{25842{,}1}{0{,}01}} = \left[-907.55,\ 2307.55\right]$$

Para muestreo sin reposición, para estimar la varianza podríamos haber tomado el estimador de Yates y Grundy:

$$\hat{V}(\hat{A}_{HT}) = \sum_{i<j}^{20} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{A_i}{\pi_i} - \frac{A_j}{\pi_j}\right)^2 = \frac{0{,}02^2 - 0{,}00038}{0{,}00038 \cdot 0{,}02^2}\sum_{i<j}^{20}\left(A_i - A_j\right)^2 =$$

$$= 131{,}58\left[9 \cdot 8(0-1)^2 + 9 \cdot 3(0-2)^2 \cdots + 8 \cdot 3(1-2)^2\right] = 26842{,}3$$

Se observa que el estimador de Yates y Grundy sobreestima la varianza en este caso.

Cuando el muestreo es con reposición, el total de microordenadores defectuosos puede estimarse mediante el estimador de Hansen y Hurwitz. Además, al ser el muestreo con probabilidades iguales, $P_i = 1/N$ y tendremos:

$$\hat{A}_{HH} = \sum_i^n \frac{A_i}{nP_i} = \sum_i^n \frac{A_i}{n\frac{1}{N}} = \frac{N}{n}\sum_i^n A_i = \frac{1000}{20}(9\cdot 0 + 8\cdot 1 + 3\cdot 2) = 700$$

La varianza se estima de la siguiente forma:

$$\hat{V}(\hat{A}) = \frac{\sum_i^n \left(\frac{A_i}{P_i} - \hat{A}\right)^2}{n(n-1)} = \frac{\sum_i^n \left(\frac{A_i}{1/N} - 700\right)^2}{n(n-1)} = \frac{\sum_{i=1}^{20} (1000\,A_i - 700)^2}{20(20-1)} = \frac{100^2 \sum_{i=1}^{20} (10\,A_i - 7)^2}{380} =$$

$$= \frac{1000}{38}\left[9(10\cdot 0 - 7)^2 + 8(10\cdot 1 - 7)^2 + 3(10\cdot 2 - 7)^2\right] = 26842,1$$

$$\hat{Cv}(\hat{A}) = \frac{\sqrt{\hat{V}(\hat{A})}}{\hat{A}} = \frac{\sqrt{26842,1}}{700} = 0,234 \quad (23,4\%)$$

La estimación por intervalos suponiendo normalidad en la población es:

$$\hat{A} \pm \lambda_\alpha \hat{\sigma}(\hat{A}) = 700 \pm 2.57\sqrt{26842,1} = [283.2, 1116.8]$$

La estimación por intervalos sin normalidad en la población es:

$$\hat{A} \pm \frac{\hat{\sigma}(\hat{A})}{\sqrt{\alpha}} = 700 \pm \sqrt{\frac{26842,1}{0,01}} = [-921.9, 2321.9]$$

Las operaciones anteriores totalmente desarrolladas se muestran a continuación.

$$\hat{Cv}(\hat{A}) = \frac{\sqrt{\hat{V}(\hat{A})}}{\hat{A}} = \frac{\sqrt{26842,3}}{700} = 0,234 \quad (23,4\%)$$

La estimación por intervalos suponiendo normalidad en la población es:

$$\hat{A} \pm \lambda_\alpha \hat{\sigma}(\hat{A}) = 700 \pm 2.57\sqrt{26842,1} = [279, 1121]$$

La estimación por intervalos sin normalidad en la población es:

$$\hat{A} \pm \frac{\hat{\sigma}(\hat{A})}{\sqrt{\alpha}} = 700 \pm \sqrt{\frac{26842,1}{0,01}} = [-938.35, 2338.35]$$

Se observa que los errores de muestreo estimados son ligeramente superiores en muestreo con reposición. Además, como es natural, los intervalos de confianza son más anchos (o sea, peores) en muestreo con reposición. La ganancia en precisión es (26842,1/25842,1-1)100=3,8%, que es una cantidad pequeña. También se observa que el estimador de Yates y Grundy para muestreo sin reposición sobreestima la varianza hasta hacerla incluso mayor que en el caso de con reposición (debido a la baja ganancia en precisión del muestreo sin reposición).

2.12. Generar una muestra de tamaño 50 de cada una de las siguientes distribuciones:

a) Uniforme entre 10 y 20
b) Poisson con λ=1

Calcular la media aritmética en cada muestra y realizar un histograma para sus valores comentando los resultados.

Para obtener muestras aleatorias según una distribución dada es necesario utilizar una herramienta adecuada. Antiguamente se usaban tablas de números aleatorios, pero en la actualidad cualquier software estadístico dispone de esta funcionalidad. Por ejemplo, Excel dispone de dos funciones para selección de números aleatorios uniformemente con reposición. La función ***ALEATORIO( )*** devuelve un número aleatorio mayor o igual que 0 y menor que 1, distribuido uniformemente. Cada vez que se calcula la hoja de cálculo, se devuelve un número aleatorio nuevo. Si desea usar ALEATORIO para generar un número aleatorio, pero no desea que los números cambien cada vez que se calcule la celda, puede escribir *=ALEATORIO( )* en la barra de fórmulas y, después, pulsar la tecla F9 para cambiar la fórmula a un número aleatorio. Para generar un número real aleatorio entre a y b, use: *ALEATORIO( )*(b-a)+a.* No obstante, la función ***ALEATORIO.ENTRE(a,b)*** devuelve un número entero aleatorio uniforme entre los números *a* y *b*

Por otra parte, Excel permite obtener números aleatorios independientes extraídos según una distribución dada utilizando herramientas de análisis. Si en el cuadro de diálogo *Análisis de datos* de la Figura 2-1 elegimos *Generación de números aleatorios*, se obtiene el cuadro de diálogo *Generación de números aleatorios* de la Figura 2-2. En el cuadro *Números de variables* introduzca el número de columnas de valores que desee incluir en la tabla de resultados. Si no introduce ningún número, Microsoft Excel rellenará todas las columnas del rango de salida que se haya especificado. En el cuadro *Cantidad de números aleatorios* introduzca el número de puntos de datos que desee ver. Cada punto de datos aparecerá en una fila de la tabla de resultados. Si no introduce ningún número, Microsoft Excel rellenará todas las columnas del rango de salida que se haya especificado. En el cuadro *Distribución* haga clic en la distribución estadística que desee utilizar para crear los valores aleatorios.

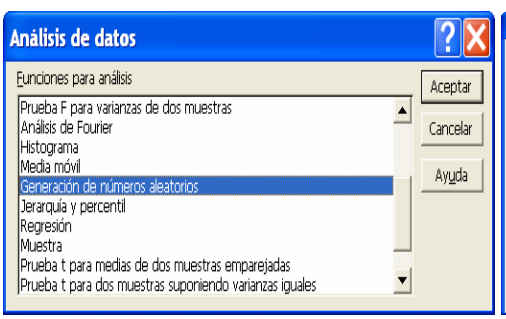


Figura 2-1

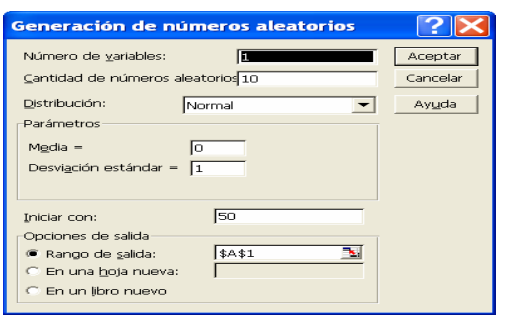


Figura 2-2

Las distribuciones posibles son:

Uniforme: Caracterizada por los límites inferior y superior. Se extraen las variables con probabilidades iguales de todos los valores del rango. Una aplicación normal utilizará una distribución uniforme en el rango 0...1.

Normal: Caracterizada por una media y una desviación estándar. Una aplicación normal utilizará una media de 0 y una desviación estándar de 1 para la distribución estándar normal.

Bernoulli: Caracterizada por la probabilidad de éxito (valor *p*) en un ensayo dado. La variables aleatorias de Bernoulli tienen el valor 0 o 1; por ejemplo, puede trazarse una variable aleatoria uniforme en el rango 0...1. Si la variable es menor o igual que la probabilidad de éxito, se asignará el valor 1 a la variable aleatoria de Bernoulli; en caso contrario, se le asignará el valor 0.

Binomial: Caracterizada por una probabilidad de éxito (valor *p*) durante un número de pruebas; por ejemplo, se pueden generar variables aleatorias Bernoulli de número de pruebas, cuya suma será una variable aleatoria binomial.

Poisson: Caracterizada por un valor *lambda*, igual a *1/media*. La distribución de Poisson se utiliza con frecuencia para caracterizar el número de incidencias por unidad de tiempo; por ejemplo, el ritmo promedio al que llegan los vehículos a una garita de peaje.

Frecuencia relativa: Caracterizada por un límite inferior y superior, un incremento, un porcentaje de repetición para valores y un ritmo de repetición de la secuencia.

Discreta: Caracterizada por un valor y el rango de probabilidades asociado. El rango debe contener dos columnas. La columna izquierda deberá contener valores y la derecha probabilidades asociadas con el valor de esa fila. La suma de las probabilidades deberá ser 1.

En el campo *Parámetros* introduzca un valor o valores para caracterizar la distribución seleccionada. En el campo *Iniciar con* escriba un valor opcional a partir del cual se generarán números aleatorios. Podrá volver a utilizar este valor para generar los mismos números aleatorios más adelante. En el cuadro *Rango de salida* introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Microsoft Excel determinará el tamaño del área de resultados y mostrará un mensaje si la tabla de resultados reemplaza datos ya existentes. Haga clic en la opción *En una hoja nueva* para insertar una hoja nueva en el libro actual y pegar los resultados comenzando por la celda A1 de la nueva hoja de cálculo. Para asignar un nombre a la nueva hoja de cálculo, escríbalo en el cuadro. Haga clic en la opción *En un libro nuevo* para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado. En la Figura 2-3 se muestra la salida correspondiente a las opciones de *Generación de números aleatorios* de la Figura 2-2 (10 números aleatorios normales de media cero y varianza 1 con semilla 50).

	A
1	-2,5043119
2	0,34869686
3	1,3207341
4	0,81364988
5	-2,3642497
6	-0,3806656
7	-2,6107955
8	0,04671847
9	0,03416289
10	0,13331032

Figura 2-3

Adicionalmente, Excel permite obtener una muestra aleatoria simple con reposición de una población numérica dada como rango de entrada. Si en el cuadro de diálogo *Análisis de datos* de la Figura 2-4 elegimos *Muestra*, se obtiene el cuadro de diálogo *Muestra* de la Figura 2-5. A continuación se explica la funcionalidad de todos los campos del cuadro de diálogo *Muestra*.

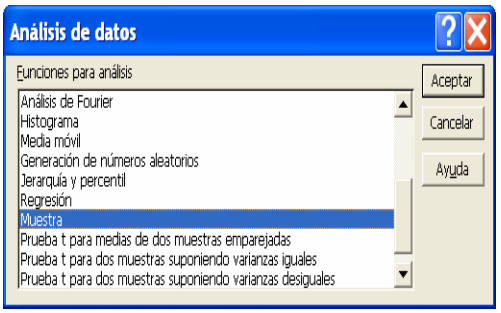


Figura 2-4

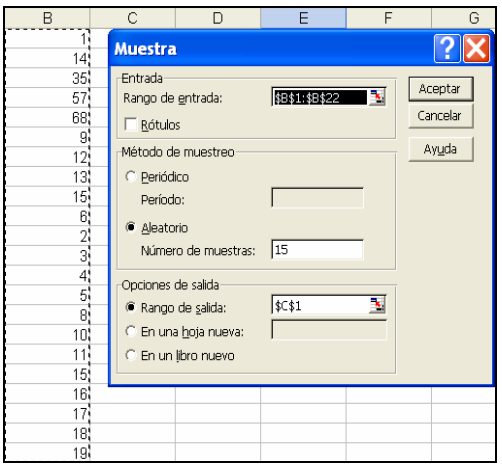


Figura 2-5

	A	B	C	D
1	-2,5043119	1	9	
2	0,34869686	14	11	
3	1,3207341	35	9	
4	0,81364988	57	57	
5	-2,3642497	68	19	
6	-0,3806656	9	15	
7	-2,6107955	12	68	
8	0,04671847	13	17	
9	0,03416289	15	10	
10	0,13331032	6	12	
11		2	1	
12		3	6	
13		4	57	
14		5	12	
15		8	9	
16		10		
17		11		
18		15		
19		16		
20		17		
21		18		
22		19		

Figura 2-6

Rango de entrada: Introduzca la referencia correspondiente al rango de datos que contenga la población de valores de los que desee extraer una muestra. Microsoft Excel extraerá muestras de la primera columna, luego de la segunda y así sucesivamente.

Rótulos: Active esta casilla si la primera fila y la primera columna del rango de entrada contienen rótulos. Desactívela si el rango de entrada carece de rótulos; Excel generará los rótulos de datos correspondientes para la tabla de resultados.

Método de muestreo: Haga clic en *Periódico* o *Aleatorio* para indicar el intervalo de muestreo que desee.

Período: Introduzca el intervalo periódico en el que desee realizar la muestra. El valor *n* del período del rango de entrada y cada valor *n* del período siguiente se copiarán en la columna de resultados. El muestreo terminará cuando se llegue al final del rango de entrada.

Número de muestras: Introduzca el número de valores aleatorios que desee en la columna de resultados. Cada valor se extrae de una posición aleatoria del rango de entrada y puede seleccionarse cualquier número más de una vez.

Rango de salida: Introduzca la referencia correspondiente a la celda superior izquierda de la tabla de resultados. Los datos se escribirán en una sola columna debajo de la celda. Si selecciona *Periódico*, el número de valores de la tabla de resultados será igual al número de valores del rango de entrada, dividido por la tasa de muestreo. Si selecciona *Aleatorio,* el número de valores de la tabla de resultados será igual al número de muestras.

En una hoja nueva: Haga clic en esta opción para insertar una hoja nueva en el libro actual y pegar los resultados comenzando por la celda A1 de la nueva hoja de cálculo. Para darle un nombre a la nueva hoja de cálculo, escríbalo en el cuadro.

En un libro nuevo: Haga clic en esta opción para crear un nuevo libro y pegar los resultados en una hoja nueva del libro creado.

Al pulsar *Aceptar* en la Figura 2-5, se obtiene la muestra aleatoria simple de tamaño 10 con reposición de la columna C de la Figura 2-6, que ha sido extraída de la población de 22 elementos de la columna B. Si la muestra se quiere sin reposición, se utiliza este mismo procedimiento hasta obtener tantos elementos distintos como tamaño muestral se requiera.

Centrándonos ya en nuestro problema particular, seleccionaremos nuestra primera muestra de tamaño 50 aleatoria uniforme de valores entre 10 y 20. Para ello, situamos la función ALEATORIO( )*(20-10)+10 en una casilla de Excel y arrastramos esta fórmula 50 casillas hacia abajo. Para seleccionar la muestra de Poisson, en *Herramientas* → *Análisis de datos* elegimos *Generación de números aleatorios* y rellenamos la pantalla de entrada como se indica en la Figura 2-7. Al pulsar *Aceptar s*e obtiene la columna de 50 números aleatorios de Poisson con $\lambda = 2$. Con las funciones PROMEDIO(A2:A51) y PROMEDIO(B2:B51) calculamos las medias de ambas columnas de números aleatorios obteniendo como resultado números cercanos a 15 y 2, que son el centro del intervalo en la distribución uniforme y el parámetro de la distribución de Poisson, respectivamente.

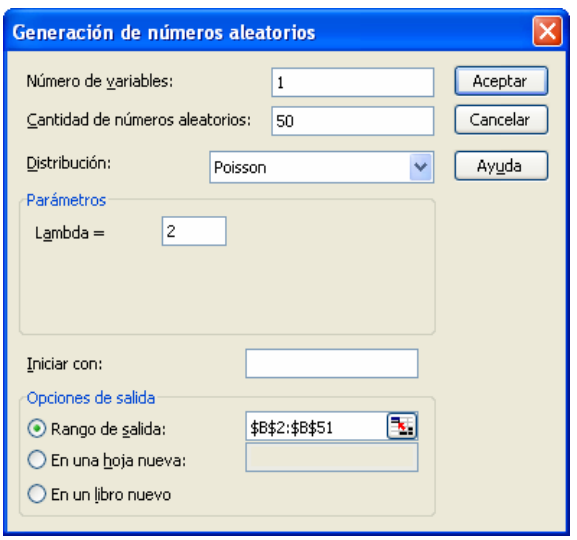


Figura 2-7

Para representar los histograma de frecuencias de cada muestra, en *Herramientas* → *Análisis de datos* (Figura 2-8) elegimos *Histograma* y rellenamos la pantalla de entrada como se indica en las Figuras 2-9 y 2-10. Al pulsar *Aceptar* se obtienen los histogramas de frecuencias. La Figura 2-11 presenta las dos series de números aleatorios con sus distribuciones de frecuencias y sus histogramas. Se observa que el histograma de la distribución de Poisson se acerca mucho a una normal.

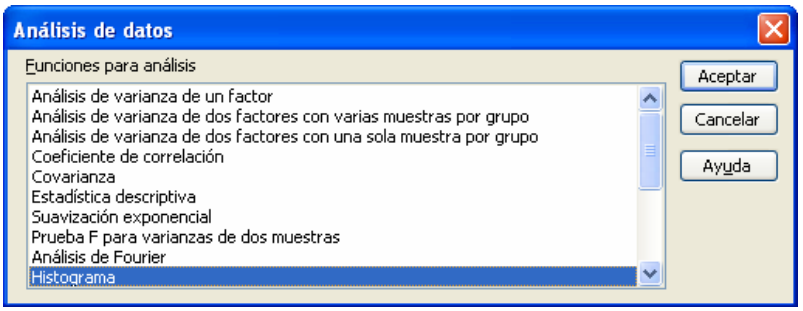


Figura 2-8

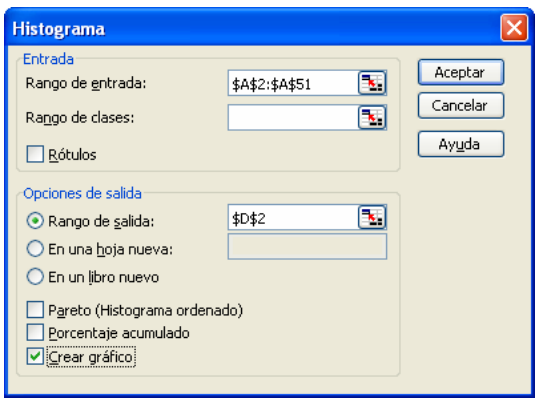


Figura 2-9

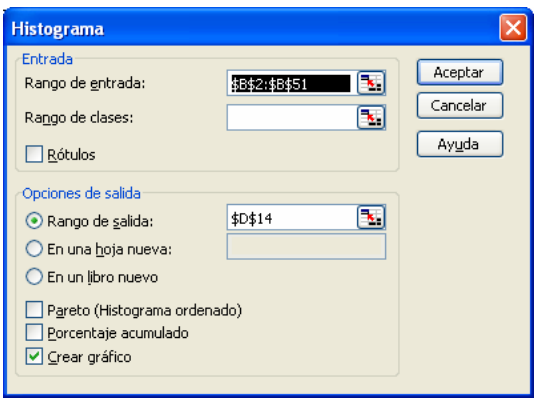


Figura 2-10

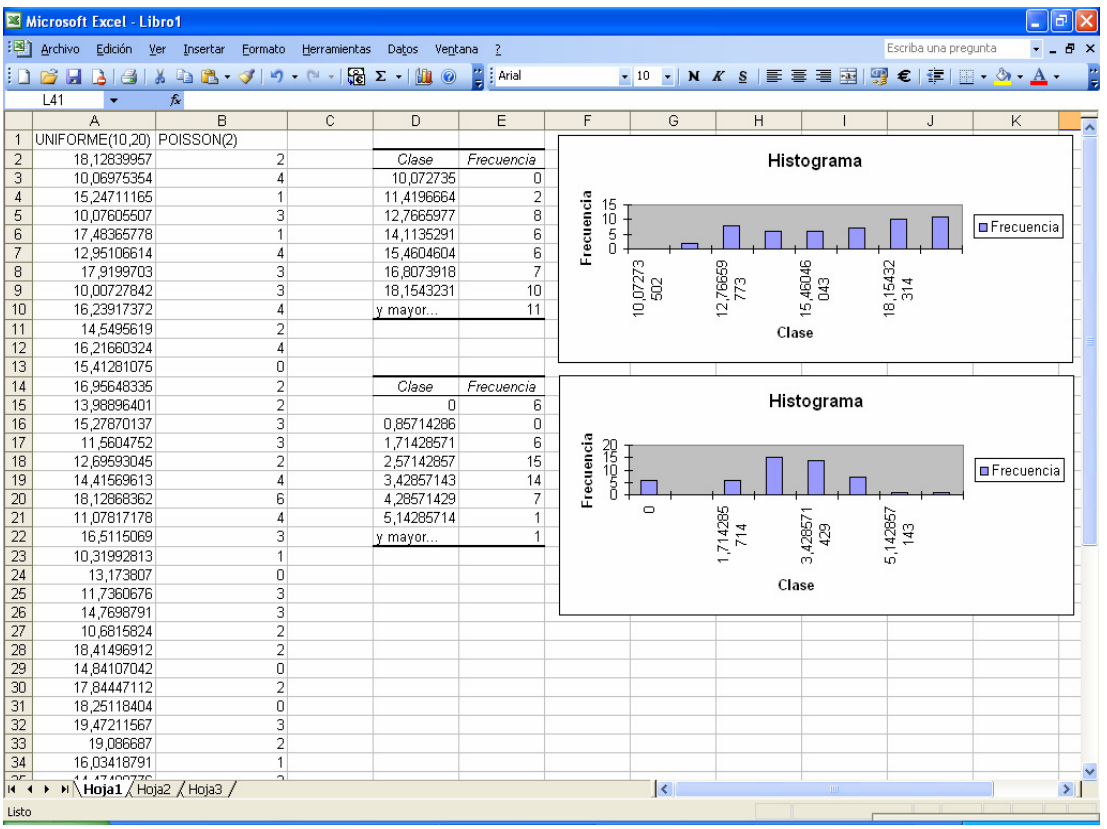


Figura 2-11

2.13. Sea la población $\{U_1, U_2, U_3\}$ en la que se conocen los valores de una determinada variable X: $X(U_1)=2$ $X(U_2)=3$ y $X(U_3)=6$. Se seleccionan dos unidades sin reemplazamiento con probabilidades proporcionales a los valores de la variable X en cada extracción, resultando elegidas las unidades U_1 y U_3. Se pide:

1) Calcular la estimación puntual lineal insesgada para el total de la variable X.
2) Calcular la estimación por intervalos al 95% para el total de la variable X (población normal).

Como el muestreo es con probabilidades proporcionales a los números 2, 3 y 6, tenemos que las probabilidades iniciales de selección de cada unidad poblacional para la muestra son $P_i = M_i/\sum M_i$, es decir: 2/11, 3/11 y 6/11. Como el método es sin reposición tomamos como estimador del total el estimador de Horwitz y Thompson y tenemos:

$$\pi_i = P_i\left(\frac{1-2P_i}{1-P_i} + \sum_{i=1}^{3}\frac{P_i}{1-P_i}\right)$$

$$\pi_1 = (2/11)\left(\frac{1-2(2/11)}{1-2/11} + \frac{2/11}{1-2/11} + \frac{3/11}{1-3/11} + \frac{6/11}{1-6/11}\right) = 0{,}468$$

$$\pi_2 = (3/11)\left(\frac{1-2(3/11)}{1-3/11} + \frac{2/11}{1-2/11} + \frac{3/11}{1-3/11} + \frac{6/11}{1-6/11}\right) = 0{,}660$$

$$\pi_3 = (6/11)\left(\frac{1-2(6/11)}{1-6/11} + \frac{2/11}{1-2/11} + \frac{3/11}{1-3/11} + \frac{6/11}{1-6/11}\right) = 0{,}871$$

$$\hat{X}_{HT} = \sum_{i=1}^{2}\frac{X_i}{\pi_i} = \frac{2}{0{,}468} + \frac{6}{0{,}871} = 11{,}16$$

Para estimar la varianza necesitamos el valor de π_{12}. Tenemos:

$\pi_{12} = P(U_1U_3)=P(U_1)P(U_3/U_1)+P(U_3)P(U_1/U_3)=(2/11)(6/9)+(6/11)(2/5)=0{,}34$

El valor anterior puede calculase también mediante:

$$\pi_{ij} = P_iP_j\left(\frac{1}{1-P_i} + \frac{1}{1-P_j}\right) = \frac{2}{11}\frac{6}{11}\left(\frac{1}{1-2/11} + \frac{1}{1-6/11}\right) = 0{,}34$$

$$\hat{V}\left(\hat{X}_{HT}\right) = \sum_{i=1}^{2}\frac{X_i^2}{\pi_i^2}(1-\pi_i) + 2\sum_{i=1}^{2}\sum_{j>i}^{2}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right) = \frac{4(1-0{,}468)}{0{,}468^2} + \frac{36(1-0{,}871)}{0{,}871^2} +$$

$$+ 2\frac{2}{0{,}468}\cdot\frac{6}{0{,}871}\cdot\frac{0{,}34-(0{,}468)(0{,}871)}{0{,}34} = 15{,}837 - 11{,}711 = 4{,}126$$

El error relativo de muestreo será $\dfrac{\sigma(\hat{X}_{HT})}{\hat{X}_{HT}}\cdot 100 = \dfrac{\sqrt{4{,}126}}{22}\cdot 100 \rightarrow 18{,}2\%$

La estimación por intervalos suponiendo normalidad en la población es:

$$\hat{X} \pm \lambda_\alpha\hat{\sigma}(\hat{X}) = 11{,}16 \pm 1{,}96\sqrt{4{,}126} = [7.17, 15.14]$$

2.14. Consideremos una región con $N = 3$ municipios con una población de 3, 5 y 7 miles de habitantes cada uno. Sabemos que la variable X = Número de mujeres en cada municipio toma los valores 1, 3, y 4 (en miles). Para estudiar el número medio de mujeres en la región se toman muestras de dos municipios con probabilidades proporcionales a sus tamaños sin reposición y sin tener en cuenta el orden de colocación de sus elementos utilizando el método de Brewer. A partir de las distribuciones en el muestreo de $\hat{X}_{HT}$ y $\hat{V}(\hat{X}_{HT})$, hallar $V(\hat{X}_{HT})$, $E(\hat{X}_{HT})$ y $E(\hat{V}(\hat{X}_{HT}))$. Comentar los resultados.

Como estamos ante un método de selección de unidades primarias compuestas con probabilidades iniciales proporcionales a los tamaños 3, 5 y 7, dichas probabilidades serán {3/15, 5/15, 7/15}. Como no hay reposición y las probabilidades son desiguales, utilizamos el estimador de Horwitz y Thompson.

Dado que el método de selección es el de Brewer tenemos:

$$\pi_i = nP_i = 2P_i \ , \ \pi_{ij} = \frac{2P_iP_j}{1+\sum_{i=1}^{N}\frac{P_i}{1-2P_i}} * \left[\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right]$$

Dado que el método es sin reposición y no importa el orden de colocación de los elementos en las muestras, el espacio muestral está constituido por la muestras (u_1,u_2), (u_1,u_3) y (u_2,u_3) con $P_1 = p(u_1) = 3/15$, $P_2 = p(u_2) = 5/15$ y $P_3 = p(u_3) = 7/15$. La distribución en el muestreo (con el esquema de selección de Brewer) del estimador de Horvitz y Thompson y del estimador de su varianza, así como el espacio muestral y las probabilidades asociadas a las muestras se presentan en el siguiente cuadro:

X_1	X_2	π_i	π_{ij}	$\hat{X}_{HT} = \frac{X_1}{2P_1}+\frac{X_2}{2P_2}$	$\hat{V}_{YG}(\hat{X}_{HT}) = \frac{\pi_1\pi_2-\pi_{12}}{\pi_{12}}\left(\frac{X_1}{\pi_1}+\frac{X_2}{\pi_2}\right)^2$
1	3	$\frac{6}{15}$	$\frac{1}{15}$	7	12
1	4	$\frac{10}{15}$	$\frac{5}{15}$	$\frac{95}{14}$	0,38265
3	4	$\frac{14}{15}$	$\frac{9}{15}$	$\frac{123}{14}$	0,00170

A partir de las distribuciones de $\hat{X}_{HT}$ y $\hat{V}(\hat{X}_{HT})$ podemos calcular su esperanza y su varianza de la siguiente forma:

$E(\hat{X}_{HT}) = 7(1/15)+(95/14)(5/15)+(123/14)(9/15) = 8$

$V(\hat{X}_{HT}) = (7-8)^2(1/15-8)^2+(95/14-8)^2(5/15-8)^2+(123/14-8)^2(9/15) = 0{,}9285$

$E(\hat{V}(\hat{X}_{HT})) = 12(1/15)+0{,}38265(5/15)+0{,}0017(9/15) = 0{,}9285$

$V(\hat{V}(\hat{X}_{HT})) = (12-0{,}9285)^2(1/15)+(0{,}38265-0{,}9285)^2(5/15)+(0{,}0017-0{,}9285)^2(9/15) = 8{,}768$

Según el resultado anterior se tiene $E(\hat{X}_{HT}) = 8 = X$, con lo que se comprueba que el estimador de Horwitz y Thompson es insesgado. También se tiene que $V(\hat{X}_{HT}) = 0.9285$ y $E(\hat{V}(\hat{X}_{HT})) = 0{,}9285 = V(\hat{X}_{HT})$, con lo que se comprueba que el estimador de la varianza es insesgado.

2.15. Resolver el problema anterior considerando ahora el esquema de selección de probabilidades gradualmente variables de Sánchez Crespo y Gabeiras con probabilidades iniciales de selección de las unidades {1/6, 1/3, 1/2}. Comparar los resultados con los obtenidos utilizando muestreo con reposición sin tener en cuenta el orden de colocación de los elementos en las muestras.

Según el esquema de probabilidades gradualmente variables, se puede suponer que existen seis bolas en una urna de las que una bola representa a la unidad u_1, dos bolas representan a la unidad u_2 y tres bolas representan a la unidad u_3, ya que $P_1 = p(u_1) = 1/6$, $P_2 = p(u_2) = 1/3 = 2/6$ y $P_3 = p(u_3) = 1/2 = 3/6$. En cada selección se extrae una única bola que no se repone a la urna para seleccionar la siguiente bola, con lo que al seleccionar la segunda bola falta una bola de la urna. Según este esquema, el espacio muestral y las probabilidades asociadas a las muestras serán:

$S(X)$	$P(u_i,u_j) = P(u_i)P(u_j/u_i) + P(u_j)P(u_i/u_j)$
(u_1,u_2)	$\frac{1}{6}\cdot\frac{1}{5}+\frac{2}{6}\cdot\frac{1}{5}=\frac{2}{15}=0{,}13333333$
(u_1,u_3)	$\frac{1}{6}\cdot\frac{3}{5}+\frac{3}{6}\cdot\frac{1}{5}=\frac{3}{15}=0{,}2$
(u_2,u_2)	$\frac{2}{6}\cdot\frac{1}{5}=\frac{1}{15}=0{,}06666666$
(u_2,u_3)	$\frac{2}{6}\cdot\frac{3}{5}+\frac{3}{6}\cdot\frac{2}{5}=\frac{6}{15}=0{,}4$
(u_3,u_3)	$\frac{3}{6}\cdot\frac{2}{5}=\frac{3}{15}=0{,}2$

El estimador insesgado para el total de Sánchez Crespo y Gabeiras es:

$$\hat{X}_{SCG}=\sum_{i=1}^{n}\frac{X_i}{nP_i}=\frac{X_1}{2P_1}+\frac{X_2}{2P_2}$$

Su varianza es $V\left(\hat{X}_{HT}\right)=\frac{M-n}{M-1}\frac{1}{n}\left(\sum_{i=1}^{n}\frac{X_i^2}{P_i}-X^2\right)=\frac{6-2}{6-1}\frac{1}{2}\left(\frac{X_1^2}{P_1}+\frac{X_2^2}{P_2}+\frac{X_3^2}{P_3}-8^2\right)$

El estimador insesgado de la varianza vale:

$$\hat{V}(\hat{X}_{SCG})=\frac{M-n}{M}\frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{X_i}{P_i}\right)^2-n\hat{X}_{SCG}^2\right]=\frac{6-2}{6}\frac{1}{2(2-1)}\left[\left(\frac{X_1}{P_1}\right)^2+\left(\frac{X_2}{P_2}\right)^2-2\hat{X}_{SCG}^2\right]$$

El cuadro del diseño muestral completo sería el siguiente:

X_1	X_2		π_{ij}	$\hat{X}_{SCG}=\frac{X_1}{2P_1}+\frac{X_2}{2P_2}$	$\hat{V}(\hat{X}_{SCG})=\frac{1}{3}\left[\left(\frac{X_1}{P_1}\right)^2+\left(\frac{X_2}{P_2}\right)^2-2\hat{X}_{SCG}^2\right]$
1	3		0,1333	7,5	1,5
1	4		0,2	7	0,6666
3	3		0,0666	9	0
3	4		0,4	8,5	0,1666
4	4		0,2	8	0

A partir del diseño anterior se tiene $E(\hat{X}_{SCG}) = (7{,}5)0{,}1333 + \ldots + 8(0{,}2) = 8 = X = 1 + 3 + 4$, con lo que se comprueba que el estimador de Sánchez Crespo y Gabeiras es insesgado. También se tiene a partir del diseño que $V(\hat{X}_{SCG}) = (7{,}5\text{-}8)^2(0{,}1333) + \ldots + (8\text{-}8)^2(0{,}2) = 0.4$ y $E(\hat{V}(\hat{X}_{SCG})) = (1{,}5)0{,}1333 + \ldots + 0(0{,}2) = 0{,}4 = V(\hat{X}_{SCG})$, con lo que el estimador de la varianza es insesgado. Por último se tiene $V(\hat{V}(\hat{X}_{SCG})) = (1{,}5-0{,}4)^2\ (0{,}1333) + \ldots +(0-0{,}4)^2(0{,}2) = 0{,}24$.

El cálculo de la varianza del estimador del total de Sánchez Crespo y Gabeiras también puede realizarse a través de su fórmula correspondiente como sigue:

$$V\left(\hat{X}_{SCG}\right) = \frac{6-2}{6-1}\frac{1}{2}\left(\sum_{i=1}^{3}\frac{X_i^2}{P_i} - X^2\right) = \frac{4}{5}\frac{1}{2}\left(\frac{X_1^2}{P_1} + \frac{X_2^2}{P_2} + \frac{X_3^2}{P_3} - 8^2\right) = \frac{4}{5}\frac{1}{2}\left(\frac{1^2}{1/6} + \frac{3^2}{1/3} + \frac{4^2}{1/2} - 8^2\right) = 0{,}4$$

Para el caso de muestreo con reposición sin importar el orden de colocación de los elementos en las muestras la probabilidad de cualquier muestra será:

$$\mathrm{P}(u_i,u_j) = \mathrm{P}(u_i)\mathrm{P}(u_j) + \mathrm{P}(u_j)\mathrm{P}(u_i) = 2\,\mathrm{P}(u_i)\mathrm{P}(u_j) \text{ y } \mathrm{P}(u_i,u_i) = [\mathrm{P}(u_i)]^2$$

Las muestras posibles son ($u1$,$u1$), ($u1$,$u2$), ($u1$,$u3$), ($u2$,$u2$) ($u2$,$u3$) y ($u3$,$u3$) con $P_1 = p(u_1) = 1/6$, $P_2 = p(u_2) = 1/3$ y $P_3 = p(u_3) = 1/2$,. Como estamos en muestreo con reposición el estimador lineal insesgado para el total es el estimador de Hansen y Hurwitz ($\hat{X}_{HH} = X_1/2P_1 + X_2/2P_2$). Como estimador insesgado para la varianza se puede utilizar:

$$\hat{V}(\hat{X}_{HH}) = \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{X_i}{P_i}\right)^2 - n\hat{X}_{HH}^2\right] = \frac{1}{2(2-1)}\left[\left(\frac{X_1}{P_1}\right)^2 + \left(\frac{X_2}{P_2}\right)^2 - 2\hat{X}_{HH}^2\right]$$

La distribución en el muestreo del estimador de Hansen y Hurwitz y del estimador de su varianza, así como el espacio muestral y las probabilidades asociadas a las muestras se presentan a continuación:

X_1	X_2		$P_{ij} = P(u_i, u_j)$	$\hat{X}_{HH} = \frac{X_1}{2P_1} + \frac{X_2}{2P_2}$	$\hat{V}(\hat{X}_{HH}) = \frac{1}{2}\left[\left(\frac{X_1}{P_1}\right)^2 + \left(\frac{X_2}{P_2}\right)^2 - 2\hat{X}_{HH}^2\right]$
1	1		0,1666	6	0
1	3		0,1666	7,5	2,25
1	4		0,1666	7	1
3	3		0,3333	9	0
3	4		0,3333	8,5	0,25
4	4		0,5	8	0

Según la tabla anterior, $E(\hat{X}_{HH}) = 6(0{,}1666) + \ldots + 8(0{,}5) = 8 = X = 1 + 3 + 4$, con lo que se comprueba que el estimador de Hansen y Hurwitz es insesgado. También se tiene que $V(\hat{X}_{HH}) = (6\text{-}8)^2(0{,}1666 + \ldots + (8\text{-}8)^2(0{,}5) = 0{,}5$ y $E(\hat{V}(\hat{X}_{HH})) = 0(0{,}1666) + \ldots +0(0{,}5) = 0{,}5 = V(\hat{X}_{HH})$, con lo que el estimador de la varianza es insesgado. Por último se tiene que $V(\hat{V}(\hat{X}_{HH})) = (0\text{-}0{,}5)^2(0{,}1666) + \ldots + (0\text{-}0{,}5)^2(0{,}5) = 0{,}5$.

El cálculo de la varianza del estimador del total de Hansen y Hurwitz también puede realizarse a través de su fórmula correspondiente como sigue:

$$V\left(\hat{X}_{HT}\right)=\frac{1}{2}\left(\sum_{i=1}^{3}\frac{X_i^2}{P_i}-X^2\right)=\frac{1}{2}\left(\frac{X_1^2}{P_1}+\frac{X_2^2}{P_2}+\frac{X_3^2}{P_3}-8^2\right)=\frac{1}{2}\left(\frac{1^2}{1/6}+\frac{3^2}{1/3}+\frac{4^2}{1/2}-8^2\right)=0{,}5$$

Observando los resultados vemos que se cumple $V(\hat{X}_{SCG})=\frac{M-n}{M-1}\cdot V(\hat{X}_{HH})$, ya que 0,4=[(6-2)/(6-1)]0,5.

Además, $\hat{V}(\hat{X}_{SCG})=\frac{M-n}{M}\cdot\hat{V}(\hat{X}_{HH})$, ya que $\hat{V}(\hat{X}_{SCG})=$ [(6-2)/6] $\hat{V}(\hat{X}_{HH})$ para todos los elementos correspondientes de las columnas consideradas en las tablas anteriores.

Como $V(\hat{X}_{SCG})$=0,4 y $V\left(\hat{X}_{HT}\right)$=0,5, el método de selección con probabilidades gradualmente variables con el estimador de Sánchez Crespo y Gabeiras resulta más preciso que el método de selección con reposición de Hansen y Hurwitz.

2.16.

Supongamos que tenemos una población de N = 5 niños para los que sus edades correspondientes en años son {3, 3, 4, 6, 8} y sus pesos en kilos son {10, 16, 16, 25, 33}. Se toman muestras sin reposición de tamaño 2 de la población de niños con probabilidades proporcionales a sus pesos. Se pide:

1) Obtener un etimador lineal insesgado para la edad media de los niños basado en la muestra de mayor probabilidad, así como su error de muestreo.
2) Si consideramos la selección de la primera unidad muestral proporcional al peso y la segunda con probabilidades iguales, obtener un estimador lineal insesgado para la edad media de los niños basado en la muestra (4,8) así como su error de muestreo.

Como no se especifica nada respecto al orden de colocación de los elementos en las muestras y el muestreo es sin reposición, supondremos que el orden no interviene. Habrá entonces $\binom{5}{2}$=10 muestras posibles, que son: (3,3), (3,4), (3,6), (3,8), (3,4), (3,6), (3,8), (4,6), (4,8) y (6,8).

Las probabilidades iniciales de selección P_i proporcionales a M_1=10, M_2=16, M_3=16, M_4=25 y M_5=33 originan los siguientes valores: P_i = {M_1/M=1/10, M_2/M=4/25, M_3/M=4/25, M_4/M=1/4, M_5/M=33/100}. Las probabilidades π_{ij} se calcularán de la siguiente forma:

$$\pi_{ij} = P((u_iu_j)\in(\tilde{x})) = P(u_i\in 1^a\cap u_j\in 2^a)+P(u_j\in 1^a\cap u_i\in 2^a)$$
$$= P(u_i\in 1^a)P(u_j\in 2^a/u_i\in 1^a)+P(u_j\in 1^a)P(u_i\in 2^a/u_j\in 1^a)=$$
$$\frac{M_i}{M}\cdot\frac{M_j}{M-M_i}+\frac{M_j}{M}\cdot\frac{M_i}{M-M_j}=P_i\cdot\frac{P_j}{1-P_i}+P_j\cdot\frac{P_i}{1-P_j}=P_iP_j\left[\frac{1}{1-P_i}+\frac{1}{1-P_j}\right]$$

Y como ya conocemos las P_i, para calcular las probabilidades π_{ij} basta sustituir en la fórmula anterior. También es posible el cálculo como sigue:

π_{11} = P(3,3) = P(3∈1ª)P(3∈2ª/3∈1ª) + P(3∈1ª)P(3∈2ª/3∈1ª) = (M_1/M)(M_2/(M-$M$1)) + (M_2/M)(M_1/(M-M_2)) = (1/10)(16/90) + (4/25)(10/84) = 0,0368

π_{12} = P(3∈1ª)P(4∈2ª/3∈1ª) + P(4∈1ª)P(3∈2ª/4∈1ª) = $(M_1/M)(M_3/(M-M_1))$ + $(M_3/M)(M_1/(M-M3))$ = (1/10)(16/90)+(4/25)(10/84) = 0,0368

π_{13} = P(3,6) = P(3∈1ª)P(6∈2ª/3∈1ª) + P(6∈1ª)P(3∈2ª/6∈1ª) = $(M_1/M)(M_4/(M-M_1))$ + $(M_4/M)(M_1/(M-M_4))$ (1/10)(25/90)+(1/4)(10/75) = 0,0611

De la misma forma se obtiene π_{14}=0,0611, π_{15}=0,0859, π_{23}=0,0609, π_{24}=0,1009, π_{25} =0,1416, π_{34} =0,1009, π_{35} =0,1416 y π_{45} =0,2331

El cálculo de los π_i se realiza de la forma siguiente:

$\pi_1 = \pi_{12}+\pi_{13}+\pi_{14}+\pi_{15}$=0,0368+0,0368+0,0611+0,0859=0,22069
$\pi_2 = \pi_{12}+\pi_{23}+\pi_{24}+\pi_{25}$=0,0368+0,0609+0,1009+0,1416=0,34039
$\pi_3 = \pi_{13}+\pi_{23}+\pi_{34}+\pi_{35}$=0,0368+0,0609+0,1009+0,1416=0,34039
$\pi_4 = \pi_{14}+\pi_{24}+\pi_{34}+\pi_{45}$=0,0611+0,1009+0,1009+0,2331=0,49614
$\pi_5 = \pi_{15}+\pi_{25}+\pi_{35}+\pi_{45}$=0,0859+0,1416+0,1416+0,2331=0,60237

También pueden calcularse los π_i mediante una expresión que los haga depender solamente de los P_i, tal y como se indica a continuación.

$$\pi_i = P(u_i \in (\tilde{x})) = P(u_i \in 1^a) + P(u_i \in 2^a \cap u_{j\neq i} \in 1^a) = P(u_i \in 1^a) +$$
$$P(u_i \in 2^a / u_{j\neq i} \in 1^a) P(u_{j\neq i} \in 1^a) = P(u_i \in 1^a) + \sum_{j\neq i} P(u_i \in 2^a / u_j \in 1^a) P(u_j \in 1^a)$$
$$= P_i + \sum_{j\neq i} \frac{M_i}{M - M_j} P_j = P_i + \sum_{j\neq i} \frac{P_i}{1-P_j} P_j = P_i \left(1 + \sum_{j\neq i} \frac{P_j}{1-P_j}\right) = P_i \left(\frac{1-2P_i+P_i}{1-P_i} + \sum_{j\neq i} \frac{P_j}{1-P_j}\right)$$
$$= P_i \left(\frac{1-2P_i}{1-P_i} + \underbrace{\frac{P_i}{1-P_i} + \sum_{j\neq i} \frac{P_j}{1-P_j}}\right) = P_i \left(\frac{1-2P_i}{1-P_i} + \sum_{j=1}^{N} \frac{P_j}{1-P_j}\right) = P_i \left(\frac{1-2P_i}{1-P_i} + \sum_{i=1}^{N} \frac{P_i}{1-P_i}\right)$$

Y como ya conocemos las P_i, para calcular las probabilidades π_{ij} basta sustituir en la fórmula anterior, con lo que se obtienen los mismos resultados. El diseño muestral será el siguiente:

$S(X)$	$P(X) = \pi_{ij}$	$\hat{X}_{HT} = \sum_{i=1}^{2} \frac{X_i}{\pi_i}$	$\hat{\bar{X}}_{HT} = \frac{1}{N} \sum_{i=1}^{2} \frac{X_i}{\pi_i}$
(3,3)	0,0368	3 / 0,22069 + 3 / 0,34039 = 22,41	4,482
(3,4)	0,0368	3 / 0,22069 + 4 / 0,34039 = 25,34	5,068
(3,6)	0,0611	3 / 0,22069 + 6 / 0,49614 = 25,69	5,138
(3,8)	0,0859	3 / 0,22069 + 8 / 0,60237 = 26,87	5,374
(3,4)	0,0609	3 / 0,34039 + 4 / 0,34039 = 20,56	4,112
(3,6)	0,1009	3 / 0,34039 + 6 / 0,49614 = 20,91	4,182
(3,8)	0,1416	3 / 0,34039 + 8 / 0,60237 = 22,09	4,418
(4,6)	0,1009	4 / 0,34039 + 6 / 0,49614 = 23,84	4,768
(4,8)	0,1416	4 / 0,34039 + 8 / 0,60237 = 25,03	5,006
(6,8)	0,2331	6 / 0,49614 + 8 / 0,60237 = 25,37	5,074

Como el muestreo es sin reposición se utiliza el estimador insesgado de Horwitz y Thompson. Para el total dicho estimador basado en la muestra de mayor probabilidad, la (6,8), vale 25,37. Para la media vale 5,074. Se estima entonces que la edad media es 5 años.

Para calcular las varianzas de estos estimadores se pueden utilizar directamente las fórmulas adecuadas, o bien se puede calcular la distribución en el muestreo de los estimadores.

Para el total tenemos:

$$V\left(\hat{X}_{HT}\right)=\sum_{i=1}^{5}\frac{X_i^2}{\pi_i}\left(1-\pi_i\right)+2\sum_{i=1}^{5}\sum_{j>i}^{5}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\left(\pi_{ij}-\pi_i\pi_j\right)=\frac{X_1^2}{\pi_1}\left(1-\pi_1\right)+\cdots+\frac{X_5^2}{\pi_5}\left(1-\pi_5\right)+$$

$$+2\left(\frac{X_1}{\pi_1}\frac{X_2}{\pi_2}\left(\pi_{12}-\pi_1\pi_2\right)+\cdots+\frac{X_4}{\pi_4}\frac{X_5}{\pi_5}\left(\pi_{45}-\pi_4\pi_5\right)\right)=\frac{3^2}{0{,}22069}\left(1-0{,}22069\right)+\cdots+\frac{8^2}{0{,}60237}\left(1-0{,}60237\right)$$

$$+2\left(\frac{3}{0{,}22069}\frac{3}{0{,}34039}\left(0{,}03683-0{,}22069*0{,}34039\right)+\cdots+\frac{6}{0{,}49614}\frac{8}{0{,}60237}\left(0{,}23313-0{,}49614*0{,}60237\right)\right)$$

=4,25.

Para la media, como $V\left(\hat{X}_{HT}\right)=N^2V(\hat{\bar{X}}_{HT})\Rightarrow V(\hat{\bar{X}}_{HT})=V(\hat{X}_{HT})/25 = 4{,}25/25 = 0{,}17$.

El estimador insesgado para la varianza basado en la muestra de mayor probabilidad (6,8) será:

$$\hat{V}\left(\hat{X}_{HT}\right)=\sum_{i=1}^{2}\frac{X_i^2}{\pi_i^2}\left(1-\pi_i\right)+2\sum_{i=1}^{2}\sum_{j>i}^{2}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_{ij}}=\frac{X_1^2}{\pi_1^2}\left(1-\pi_1\right)+\frac{X_2^2}{\pi_2^2}\left(1-\pi_2\right)+2\left(\frac{X_1}{\pi_1}\frac{X_2}{\pi_2}\frac{\left(\pi_{12}-\pi_1\pi_2\right)}{\pi_{12}}\right)=0{,}41$$

Para la media, $\hat{V}(\hat{\bar{X}}_{HT})=\hat{V}(\hat{X}_{HT})/25 = 0{,}41/25 = 0{,}016$.

Para el segundo apartado del problema las probabilidades P_i proporcionales a M_1=10, M_2=16, M_3=16, M_4=25 y M_5=33 en la primera extracción tienen los siguientes valores: $P_1=M_1/M$=1/10, $P_2=M_2/M$=4/25, $P_3=M_3/M$=4/25, $P_4=M_4/M$=1/4 y $P_5=M_5/M$=33/100. Las probabilidades iguales en segunda extracción valdrán 1/4. Las probabilidades π_{ij} se calcularán de la siguiente forma:

$$\begin{aligned}\pi_{ij} &= P((u_iu_j)\in(\tilde{x}))=P(u_i\in 1^a\cap u_j\in 2^a)+P(u_j\in 1^a\cap u_i\in 2^a)\\ &= P(u_i\in 1^a)P(u_j\in 2^a/u_i\in 1^a)+P(u_j\in 1^a)P(u_i\in 2^a/u_j\in 1^a)=\\ &\frac{M_i}{M}\cdot\frac{1}{4}+\frac{M_j}{M}\cdot\frac{1}{4}=P_i\cdot\frac{1}{4}+P_j\cdot\frac{1}{4}=\frac{P_i+P_j}{4}\end{aligned}$$

Calculamos ahora los π_i mediante una expresión que los haga depender solamente de los P_i, tal y como se indica a continuación.

$$\begin{aligned}\pi_i &= P(u_i\in(\tilde{x}))=P(u_i\in 1^a)+P(u_i\in 2^a\cap u_{j\neq i}\in 1^a)\\ &= P(u_i\in 1^a)+P(u_i\in 2^a/u_{j\neq i}\in 1^a)P(u_{j\neq i}\in 1^a)\\ &= P(u_i\in 1^a)+\sum_{j\neq i}P(u_i\in 2^a/u_j\in 1^a)P(u_j\in 1^a)\\ &= P_i+\sum_{j\neq i}\frac{1}{4}P_j=P_i+\frac{1}{4}\sum_{j\neq i}P_j=P_i+\frac{1}{4}(1-P_i)=\frac{3}{4}P_i+\frac{1}{4}\end{aligned}$$

Se observa que estamos ante el ***método de selección sin reposición de Ikeda*** para el caso de tamaño de muestra n=2, con lo que las π_i y π_{ij} también podrían haberse calculado mediante las expresiones siguientes (se obtendrían los mismos resultados):

$$\pi_i = P_i + (1 - P_i) * \frac{n-1}{N-1} = \frac{N-n}{N-1} * P_i + \frac{n-1}{N-1}$$

$$\pi_{ij} = \frac{n-1}{N-1} * \left[\frac{N-n}{N-2} (P_i + P_j) + \frac{n-2}{N-2} \right]$$

Ya tenemos todos los datos para calcular los valores de π_i y π_{ij}, pues sólo dependen de P_i y P_j que son datos. También podemos calcular ya el estimador $\hat{X}_{HT}$. El diseño muestral será:

$S(X)$	$P(X) = \pi_{ij} = \frac{P_i + P_j}{4}$	$\hat{X}_{HT} = \sum_{i=1}^{2} \frac{X_i}{\pi_i}$	$\pi_i = \frac{3}{4} P_i + \frac{1}{4}$
(3,3)	0,065	3 / 0,325 + 3 / 0,37 = 17,34	
(3,4)	0,065	3 / 0,325 + 4 / 0,37 = 20,04	
(3,6)	0,0875	3 / 0,325 + 6 / 0,4375 = 22,95	0,325
(3,8)	0,1075	3 / 0,325 + 8 / 0,4975 = 25,31	0,37
(3,4)	0,08	3 / 0,37 + 4 / 0,37 = 18,92	0,37
(3,6)	0,1025	3 / 0,37 + 6 / 0,4375 = 21,82	0,4375
(3,8)	0,1225	3 / 0,37 + 8 / 0,4975 = 24,19	0,4975
(4,6)	0,1025	4 / 0,37 + 6 / 0,4375 = 24,53	
(4,8)	0,1225	4 / 0,37 + 8 / 0,4975 = 26,90	
(6,8)	0,145	6 / 0,4375 + 8 / 0,4975 = 29,8	

Vemos que para la muestra (4,8) el estimador insesgado de Horvitz y Thompson para el total poblacional vale 26,90 y para la media 26,90/5 = 5,38. Sigue obteniéndose que la edad media estimada de los niños es 5 años aproximadamente.

Para hallar la varianza del estimador del total se puede utilizar su distribución en el muestreo o bien se puede aplicar directamente la fórmula apropiada tal y como se indica a continuación:

$$V(\hat{X}_{HT}) = \sum_{i=1}^{5} \frac{X_i^2}{\pi_i}(1 - \pi_i) + 2 \sum_{i=1}^{5} \sum_{j>i}^{5} \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j) = \frac{X_1^2}{\pi_1}(1 - \pi_1) + \cdots + \frac{X_5^2}{\pi_5}(1 - \pi_5) +$$

$$+ 2 \left(\frac{X_1}{\pi_1} \frac{X_2}{\pi_2} (\pi_{12} - \pi_1 \pi_2) + \cdots + \frac{X_4}{\pi_4} \frac{X_5}{\pi_5} (\pi_{45} - \pi_4 \pi_5) \right) = \frac{3^2}{0,325}(1 - 0,325) + \cdots + \frac{8^2}{0,4975}(1 - 0,4975)$$

$$+ 2 \left(\frac{3}{0,325} \frac{3}{0,37} (0,065 - 0,325 * 0,37) + \cdots + \frac{6}{0,4375} \frac{8}{0,4975} (0,145 - 0,4375 * 0,4975) \right) = 12,66$$

El estimador insesgado para la varianza basado en la muestra (4,8) será:

$$\hat{V}\left(\hat{X}_{HT}\right)=\sum_{i=1}^{2}\frac{X_i^2}{\pi_i^2}\left(1-\pi_i\right)+2\sum_{i=1}^{2}\sum_{j>i}^{2}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_{ij}}=\frac{X_1^2}{\pi_1^2}\left(1-\pi_1\right)+\frac{X_2^2}{\pi_2^2}\left(1-\pi_2\right)+2\left(\frac{X_1}{\pi_1}\frac{X_2}{\pi_2}\frac{\left(\pi_{12}-\pi_1\pi_2\right)}{\pi_{12}}\right)$$

$$=\frac{4^2}{0{,}37^2}\left(1-0{,}37\right)+\frac{8^2}{0{,}4975^2}\left(1-0{,}4975\right)+2\left(\frac{4}{0{,}37}\frac{8}{0{,}4975}\frac{\left(0{,}1225-0{,}37*0{,}4975\right)}{0{,}1225}\right)=43{,}3$$

Para la media se tiene que $\hat{V}\left(\hat{\bar{X}}_{HT}\right)=\frac{1}{25}\hat{V}\left(\hat{X}_{HT}\right)=1{,}73$.

Para hallar el estimador insesgado para la varianza basado en la muestra (4,8) también se puede usar el estimador insesgado de Yates y Grundy de la forma siguiente:

$$\hat{V}\left(\hat{X}_{HT}\right)=\sum_{i=1}^{2}\sum_{j>i}^{2}\left(\frac{X_i}{\pi_i}-\frac{X_j}{\pi_j}\right)^2\frac{\left(\pi_i\pi_j-\pi_{ij}\right)}{\pi_{ij}}=\left(\frac{X_1}{\pi_1}-\frac{X_2}{\pi_2}\right)^2\frac{\left(\pi_1\pi_2-\pi_{12}\right)}{\pi_{12}}=\left(\frac{4}{0{,}37}-\frac{8}{0{,}4975}\right)^2\frac{\left(0{,}37*0{,}4975-0{,}1225\right)}{0{,}1225}=13{,}958$$

Para la media, $\hat{V}\left(\hat{\bar{X}}_{HT}\right)=\frac{1}{25}\hat{V}\left(\hat{X}_{HT}\right)=0{,}55$

Se observa que para la muestra (4,8) el estimador de Yates y Grundy para la varianza del total resulta más preciso que el estimador de la varianza de Horwitz y Thompson.

EJERCICIOS PROPUESTOS

2.1. Supongamos que tenemos una población de $N = 3$ unidades primarias de la que se obtienen todas las muestras posibles de tamaño $n = 2$ con probabilidades iguales y bajo los siguientes supuestos:

Muestreo sin reposición sin intervenir el orden
Muestreo sin reposición interviniendo el orden
Muestreo con reposición sin intervenir el orden
Muestreo con reposición interviniendo el orden

Se pide:

Hallar el espacio muestral asociado a los cuatro tipos de muestreo y las probabilidades asociadas a las muestras.
Si al medir una variable X sobre los elementos de la población se obtienen los valores {1, 3, 4}, ¿cuál de todos los métodos de muestreo es más preciso al estimar el total poblacional mediante un estimador lineal insesgado apropiado?

2.2. Una población consta de 40000 unidades distribuidas en 400 conglomerados de 100 unidades cada uno. Una muestra aleatoria con probabilidades iguales sin reposición de tamaño 25 conglomerados presenta los siguientes datos:

Total de unidades de la clase C	12	17	23	33	36
Nº de conglomerados de la muestra	2	3	9	5	6

Estimar el total y la proporción de unidades de la población que pertenecen a la clase C, así como sus errores de muestreo absolutos y relativos.

2.3. Supongamos que tenemos una población de $N = 5$ unidades primarias para las que una variable X medida sobre ellas proporciona los valores 3, 3, 4, 6 y 8. Se toma una muestra de tamaño $n = 2$ sin reposición asignando en la primera extracción probabilidades proporcionales a los números 10, 16, 16, 25 y 33, y también en la segunda (prescindiendo de la unidad seleccionada en primer lugar). Se pide:

Calcular las probabilidades π_{ij} $(i \neq j)$ y comprobar que $\sum \pi_i = 2$ para i = 1, 2, ..., 5

Comprobar también que $\sum_{\substack{i=1 \\ i \neq j}}^{N} \pi_i = n - \pi_j$ y $\sum_{\substack{i=1 \\ i \neq j}}^{N} \pi_{ij} = (n-1)\pi_j$.

Obtener estimadores lineales insesgados para el total y la media (para la muestra de mayor probabilidad), así como sus errores de muestreo.

2.4. Supongamos que tenemos una población de $N = 3$ unidades primarias para las que una variable X medida sobre ellas proporciona los valores {1, 3, 4} con probabilidades de selección proporcionales a los tamaños 3, 5 y 7. Se toman muestras de tamaño n=2 sin reposición y sin tener en cuenta el orden de colocación de los elementos mediante el método de selección de Durbin. A partir de las distribuciones en el muestreo de $\hat{X}_{HT}$ y $\hat{V}(\hat{X}_{HT})$, hallar $V(\hat{X}_{HT})$, $E(\hat{X}_{HT})$ y $E(\hat{V}(\hat{X}_{HT}))$. Comentar los resultados.

CAPÍTULO

3

MUESTREO ALEATORIO SIMPLE SIN Y CON REPOSICIÓN. SUBPOBLACIONES

OBJETIVOS

1. Introducir el concepto de muestreo aleatorio simple.
2. Comprender las especificaciones del muestreo aleatorio simple sin reposición o muestreo irrestricto aleatorio.
3. Analizar el muestreo aleatorio simple sin reposición.
4. Estudiar las estimaciones, errores y estimación de los errores en muestreo aleatorio simple sin reposición.
5. Especificar los factores de elevación en muestreo aleatorio simple sin reposición.
6. Evaluar el tamaño de la muestra en muestreo aleatorio simple sin reposición.
7. Comprender las especificaciones del muestreo aleatorio simple con reposición.
8. Analizar el muestreo aleatorio simple con reposición.
9. Estudiar las estimaciones, errores y estimación de los errores en muestreo aleatorio simple con reposición.
10. Especificar los factores de elevación en muestreo aleatorio simple con reposición.
11. Evaluar el tamaño de la muestra en muestreo aleatorio simple con reposición.
12. Comparar el muestreo aleatorio simple con y sin reposición.
13. Obtener estimadores en subpoblaciones con y sin reposición.
14. Calcular errores y estimación de los errores en subpoblaciones con y sin reposición.

ÍNDICE

1. Muestreo aleatorio simple sin reposición. Especificaciones

2. Estimadores, varianzas y estimación de varianzas.

3. Tamaño de la muestra.

4. Muestreo aleatorio simple con reposición. Estimadores

5. Varianzas y su estimación con reposición.

6. Tamaño de la muestra con reposición.

7. Comparación entre muestreo aleatorio sin y con reposición.

8. Subpoblaciones.

9. Problemas resueltos.

10. Ejercicios propuestos.

MUESTREO ALEATORIO SIMPLE SIN REPOSICIÓN. ESPECIFICACIONES

El muestreo aleatorio simple sin reposición es un procedimiento de selección de muestras con probabilidades iguales, que consiste en obtener la muestra unidad a unidad de forma aleatoria sin reposición a la población de las unidades previamente seleccionadas, teniendo presente que el orden de colocación de los elementos en las muestras no interviene (es decir, que muestras con los mismos elementos colocados en orden distinto se consideran iguales). De esta forma, las muestras con elementos repetidos son imposibles. Como el procedimiento de selección es con probabilidades iguales, todas las muestras son equiprobables, y además se cumple que todas las unidades de la población tienen la misma probabilidad de pertenecer a la muestra $\pi_i = n/N$. Se supone que el tamaño de la población es N y el tamaño de la muestra es n. Como la muestra se selecciona sin reposición, se realiza la selección sucesiva de las unidades para la muestra con probabilidades $1/(N-t)$ para valores de $t = 0, 1, ..., n$.

Podríamos resumir las especificaciones del muestro aleatorio simple sin reposición o ***muestreo irrestricto aleatorio*** como sigue:

- Se trata de un tipo de muestreo de unidades elementales.
- Consiste en obtener la muestra unidad a unidad de forma aleatoria sin reposición a la población de las unidades previamente seleccionadas.
- El orden de colocación de los elementos en las muestras no interviene; es decir, las muestras con los mismos elementos colocados en orden distinto se consideran iguales.
- Las muestras con elementos repetidos son imposibles.
- Se trata de un procedimiento de selección con probabilidades iguales porque todas las unidades de la población van a tener la misma probabilidad de pertenecer a la muestra.
- Todas las muestras son equiprobables.

Probabilidad de una muestra cualquiera

En la selección de una muestra aleatoria simple sin reposición de n elementos de entre los N de la población, el espacio muestral asociado tiene un número total de muestras igual a:

$$C_{N,n} = \binom{N}{n}$$

ya que el orden de colocación de los elementos en las muestras no interviene. Como el procedimiento es con probabilidades iguales, la probabilidad de una muestra cualquiera será:

$$p(u_1, \cdots, u_n) = \frac{Casos\ favorables}{Casos\ posibles} = \frac{1}{C_{N,n}} = \frac{1}{\binom{N}{n}}$$

Estamos entonces ante un procedimiento de selección con muestra equiprobables.

Probabilidad π_i que tiene una unidad de la población de pertenecer a la muestra

Para calcular la probabilidad π_i que tiene una unidad de la población de pertenecer a la muestra observamos que el número de muestras posibles de tamaño n en selección irrestricta aleatoria es:

$$C_{N,n} = \binom{N}{n}$$

Por otra parte, el número de muestras posibles que se pueden formar con los elementos de la población y que contengan al elemento dado u_i será:

$$C_{N-1,n-1} = \binom{N-1}{n-1}$$

ya que en este caso se fija el elemento u_i y las muestras posibles resultan de las formas posibles de seleccionar de entre los N-1 elementos de la población restantes n-1 de ellos para la muestra (el elemento u_i ya está fijo en la muestra).

Tenemos entonces:

$$\pi_i = P\left(u_i \in (\tilde{x})\right) = \frac{Casos\ \ favorables}{Casos\ \ posibles} =$$

$$= \frac{N^{o}\ de\ \ muestras\ \ que\ \ contienen\ \ la\ \ unidad\ \ u_i}{N^{o}\ \ total\ \ de\ muestras}$$

$$= \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N.(N-1)!}{n(n-1)!(N-n)!}} = \frac{1}{\frac{N}{n}} = \frac{n}{N}$$

Como todas las unidades de la población tienen la misma probabilidad de pertenecer a la muestra, estamos ante un procedimiento de selección con probabilidades iguales.

ESTIMADORES, VARIANZAS Y ESTIMACIÓN DE VARIANZAS

Ya sabemos que el estimador lineal insesgado general para el caso de muestreo sin reposición es el estimador de Horvitz y Thompson $\hat{\theta}_{HT}$.

Se tiene que $\hat{\theta}_{HT} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}$ estima $\theta = \sum_{i=1}^{N} Y_i$, con $E\left(\hat{\theta}\right) = \theta$, es decir, insesgadamente, siendo π_i la probabilidad de que la unidad u_i pertenezca a la muestra ($\pi_i = n/N$).

Entonces podemos deducir los estimadores lineales insesgados para el total ($Y_i=X_i$), media ($Y_i=X_i/N$), proporción ($Y_i=A_i/N$) y total de clase ($Y_i=A_i$) como sigue:

$$\theta = X = \sum_{i=1}^{N} X_i \Rightarrow Y_i = X_i \Rightarrow \hat{\theta} = \hat{X} = \sum_{i=1}^{n} \frac{X_i}{\pi_i} = \sum_{i=1}^{n} \frac{X_i}{\frac{n}{N}} = N \underbrace{\frac{1}{n} \sum_{i=1}^{n} X_i}_{\bar{x}} = N\bar{x}$$

$$\theta = \bar{X} = \sum_{i=1}^{N} \frac{X_i}{N} \Rightarrow Y_i = \frac{X_i}{N} \Rightarrow \hat{\theta} = \hat{\bar{X}} = \sum_{i=1}^{n} \frac{\frac{X_i}{N}}{\pi_i} = \sum_{i=1}^{n} \frac{\frac{X_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{x}$$

$$\theta = P = \sum_{i=1}^{N} \frac{A_i}{N} \Rightarrow Y_i = \frac{A_i}{N} \Rightarrow \hat{\theta} = \hat{P} = \sum_{i=1}^{n} \frac{\frac{A_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^{n} A_i$$

$$\theta = A = \sum_{i=1}^{N} A_i \Rightarrow Y_i = A_i \Rightarrow \hat{\theta} = \hat{A} = \sum_{i=1}^{n} \frac{A_i}{\frac{n}{N}} = N \frac{1}{n} \sum_{i=1}^{n} A_i = N\hat{P}$$

Se observa que los estimadores de la media y la proporción poblacional son los estimadores por analogía (media y proporción muestral), mientras que los estimadores del total y el total de clase poblacionales son la expansión mediante el tamaño poblacional de la media y proporción muestrales (en este caso, $\hat{X} = N\bar{x} = (N/n)x \Rightarrow$ los factores de elevación son N/n).

Varianzas de los estimadores

Sabemos que la varianza del estimador de Horvitz y Thompson está dada por la expresión:

$$V\left(\hat{\theta}_{HT}\right) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}\left(1 - \pi_i\right) + 2\sum_{i<j}^{N} \frac{Y_i Y_j}{\pi_i \pi_j}\left(\pi_{ij} - \pi_i \pi_j\right)$$

Para el caso particular del muestreo aleatorio simple sin reposición se sabe que $\pi_i = n/N$ y $\pi_{ij} = n(n-1) / [N(N-1)]$. Considerando el estimador del total y sustituyendo estos valores de π_i y π_{ij} en la expresión de la varianza tenemos:

$$V\left(\hat{X}\right) = \sum_{i=1}^{N} \frac{X_i^2}{\frac{n}{N}}\left(1 - \frac{n}{N}\right) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{X_i X_j}{\frac{n}{N}\frac{n}{N}}\left(\frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N}\right)$$

$$= N^2\left(1 - \frac{n}{N}\right)\frac{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}{n} = N^2\left(1 - f\right)\frac{S^2}{n}$$

Para los estimadores de la media, total y total de clase tenemos:

$$V\left(\hat{\bar{X}}\right) = \left(1 - f\right)\frac{S^2}{n} \qquad V\left(\hat{P}\right) = \left(1 - f\right)\frac{S^2}{n} = \left(1 - f\right)\frac{\frac{N}{N-1}PQ}{n} = \frac{N}{N-1}\frac{1}{n}\left(1 - f\right)PQ$$

$$V\left(\hat{A}\right) = N^2\left(1 - f\right)\frac{S^2}{n} = N^2\left(1 - f\right)\frac{\frac{N}{N-1}PQ}{n} = \frac{N^3}{N-1}\frac{1}{n}\left(1 - f\right)PQ$$

Estimación de varianzas

Sabemos que la varianza del estimador de Horvitz y Thompson está dada por la expresión:

$$\hat{V}\left(\hat{\theta}_{HT}\right)=\sum_{i=1}^{n}\frac{Y_i^2}{\pi_i^2}\left(1-\pi_i\right)+\sum_{i<j}^{n}\frac{Y_iY_j}{\pi_i\pi_j}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}$$

Si aplicamos la expresión anterior al estimador del total tendremos:

$$\hat{V}\left(\hat{X}\right)=\sum_{i=1}^{n}\frac{X_i^2}{\frac{n^2}{N^2}}\left(1-\frac{n}{N}\right)+2\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{X_iX_j}{\frac{n}{N}\frac{n}{N}}\frac{\left(\frac{n(n-1)}{N(N-1)}-\frac{n}{N}\frac{n}{N}\right)}{\frac{n(n-1)}{N(N-1)}}=$$

$$\frac{N(N-n)}{n}\underbrace{\frac{1}{n-1}\left[\sum_{i=1}^{n}\left(X_i-\bar{x}\right)^2\right]}_{\hat{S}^2}=N^2\frac{(N-n)}{n}\frac{\hat{S}^2}{N}=N^2(1-f)\frac{\hat{S}^2}{N}$$

Para los estimadores de la media, total y total de clase tenemos:

$$\hat{V}(\hat{\bar{X}})=(1-f)\frac{\hat{S}^2}{n}\qquad\hat{V}\left(\hat{P}\right)=(1-f)\frac{\hat{S}^2}{n}=(1-f)\frac{\frac{n}{n-1}\hat{P}\hat{Q}}{n}=(1-f)\frac{1}{n-1}\hat{P}\hat{Q}$$

$$\hat{V}\left(\hat{A}\right)=N^2(1-f)\frac{\hat{S}^2}{n}=N^2(1-f)\frac{\frac{n}{n-1}\hat{P}\hat{Q}}{n}=N^2(1-f)\frac{1}{n-1}\hat{P}\hat{Q}$$

De las fórmulas de las varianzas y sus estimaciones, se deduce que en muestreo aleatorio simple sin reposición la cuasivarianza muestral

$$\hat{S}^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{x}\right)^2$$

es un estimador insesgado de la cuasivarianza poblacional $S^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2$.

TAMAÑO DE LA MUESTRA

Estudiaremos el ***tamaño de muestra necesario para cometer un error de muestreo*** $e=\sigma(\hat{\theta})$ dependiendo de si $\hat{\theta}$ estima la media, el total, la proporción o el total de clase.

Media:

$$e=\sigma(\hat{\bar{X}})=\sqrt{(1-f)\frac{S^2}{n}}\Rightarrow e^2=\left(1-\frac{n}{N}\right)\frac{S^2}{n}=\frac{S^2}{n}-\frac{S^2}{N}$$

$$\Rightarrow\frac{S^2}{n}=e^2+\frac{S^2}{N}\Rightarrow n=\frac{S^2}{e^2+\frac{S^2}{N}}=\frac{NS^2}{Ne^2+S^2}$$

Se observa que cuando $N\to\infty$ (fracción de muestreo n/N tendiendo a cero) el tamaño muestral $n\to S^2/e^2=n_0$ (n inversamente proporcional al cuadrado del error de muestreo).

La expresión del tamaño muestral n puede ponerse en función de N y del valor n_0 como sigue:

$$n = \frac{S^2}{e^2 + \frac{S^2}{N}} = \frac{S^2/e^2}{1 + \frac{S^2/e^2}{N}} = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{n_0 N}{n_0 + N} = f(N)$$

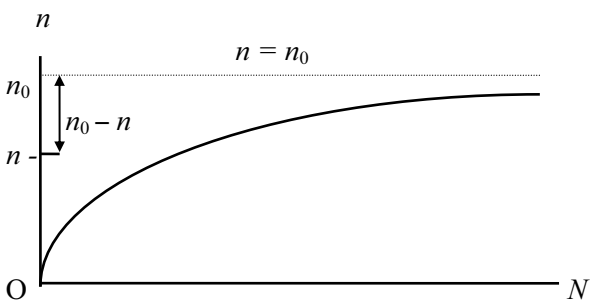
Si representamos gráficamente la curva de ecuación $n = f(N)$ observamos que pasa por el origen de coordenadas, ya que $f(0) = 0$, que tiene una asíntota paralela al eje OX de ecuación $n = n_0$, ya que $\lim_{N\to\infty} f(N) = n_0$, que es siempre creciente dado que la primera derivada:

$$f'(N) = \frac{n_0^2}{(n_0 + N)^2}$$

es siempre positiva, que no tiene máximos ni mínimos dado que la ecuación definida por $f'(N) = 0$ no tiene solución en N, que es siempre convexa ya que la segunda derivada:

$$f''(N) = -\frac{2n_0^2}{(n_0 + N)^3}$$

es siempre negativa y que no tiene puntos de inflexión ya que que la ecuación definida por $f''(N)=0$ no tiene solución en N. Por tanto, la representación gráfica de $n = f(N)$ es la siguiente:



Como la curva $n = f(N)$ es creciente, al aumentar el tamaño poblacional N también aumenta el tamaño muestral n necesario para un error de muestreo dado. Pero como n ha de ser un número entero y la curva $n=n_0$ es una asíntota horizontal, desde un cierto N en adelante los aumentos de N no producen aumentos en n. Precisamente los aumentos de N no producen aumentos en n cuando $|n_0 - n|<1$. Pero:

$$|n_0 - n| = \left| n_0 - \frac{n_0 N}{n_0 + N} \right| = \frac{n_0^2}{n_0 + N} < 1 \Rightarrow n_0^2 < n_0 + N \Rightarrow N > n_0(n_0 - 1) = \frac{S^2}{e^2}\left(\frac{S^2}{e^2} - 1\right)$$

Luego la misma precisión da una muestra de tamaño n para una población de N elementos que para una población de N ' elementos con N '> N siempre y cuando se cumpla que:

$$N > n_0(n_0 - 1) = \frac{S^2}{e^2}\left(\frac{S^2}{e^2} - 1\right)$$

Total:

$$e = \sigma(\hat{X}) = \sqrt{N^2(1-f)\frac{S^2}{n}} \Rightarrow e^2 = N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n} = \frac{N^2S^2}{n} - \frac{N^2S^2}{N} \Rightarrow$$

$$\Rightarrow \frac{N^2S^2}{n} = e^2 + \frac{N^2S^2}{N} \Rightarrow n = \frac{N^2S^2}{e^2 + \frac{N^2S^2}{N}} = \frac{N^3S^2}{\underbrace{Ne^2 + N^2S^2}_{N\left(e^2+NS^2\right)}} = \frac{N^2S^2}{e^2 + NS^2}$$

La expresión anterior también puede escribirse como:

$$n = \frac{N^2\left(\frac{S}{e}\right)^2}{1 + N\left(\frac{S}{e}\right)^2} = \frac{N^2 n_1}{1 + Nn_1} = f(N)$$

Si representamos gráficamente la curva de ecuación $n = f(N)$ observamos que pasa por el origen de coordenadas ya que $f(0) = 0$, que tiene una asíntota oblicua de ecuación $n = N - 1/n_1$ ya que:

$$\lim_{N\to\infty}\frac{f(N)}{N} = 1 \text{ y } \lim_{N\to\infty}\left(f(N) - N\right) = \lim_{N\to\infty}\frac{-N}{1+n_1N} = -\frac{1}{n_1}$$

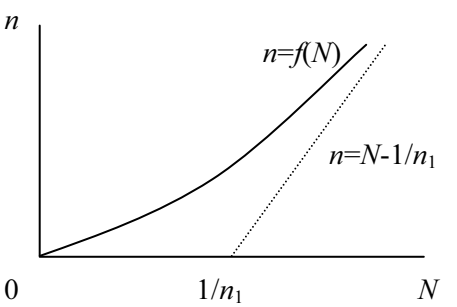
Además es siempre creciente ya que la primera derivada:

$$f'(N) = \frac{2n_1N + n_1^2N^2}{(1+n_1N)^2}$$

es siempre positiva, que no tiene máximos ni mínimos ya que la ecuación definida por $f'(N)=0$ no tiene solución en N, que es siempre cóncava puesto que:

$$f''(N) = \frac{2n_1^2N}{(1+n_1N)^3}$$

es siempre positiva. Por tanto, la representación gráfica de $n = f(N)$ es la siguiente:



Observando la gráfica de $n = f(N)$ se ve que n siempre crece al crecer N, es decir, que al aumentar el tamaño poblacional también aumentará el tamaño de muestra necesario para cometer un error de muestreo prefijado.

Proporción:

Si sustituimos el valor de S^2 para variables A_i (que sólo toman los valores 0 y 1) en la fórmula del tamaño muestral para la media tendremos para la estimación de la proporción el tamaño:

$$n = \frac{NS^2}{Ne^2 + S^2} = \frac{N\dfrac{N}{N-1}PQ}{\dfrac{N}{N-1}PQ + Ne^2} = \frac{N^2PQ}{\underbrace{NPQ + (N-1)Ne^2}_{N\left(e^2(N-1)+PQ\right)}} = \frac{NPQ}{e^2(N-1)+PQ}$$

En el caso de la proporción se observa que cuando $N \to \infty$ (fracción de muestreo n/N tendiendo a cero) el tamaño muestral $n \to S^2/e^2 = \dfrac{N}{N-1}PQ \Big/ e^2 \cong PQ/e^2 = n_0$ (n inversamente proporcional al cuadrado del error de muestreo y directamente proporcional a la proporción poblacional P). En este caso, la misma precisión da una muestra de tamaño n para una población de N elementos que para una población de N ' elementos con $N' > N$ siempre y cuando se cumpla la desigualdad definida por:

$$N > n_0(n_0 - 1) = \frac{\dfrac{N}{N-1}PQ}{e^2}\left(\frac{\dfrac{N}{N-1}PQ}{e^2} - 1\right) \cong \frac{PQ}{e^2}\left(\frac{PQ}{e^2} - 1\right)$$

Para la estimación de la proporción es muy interesante tener en cuenta que para poblaciones grandes o fracción de muestreo pequeña ($N \to \infty$), el valor máximo de n se obtiene para $P = Q = 1/2$. Para constatar este resultado sabemos que si $N \to \infty$ el tamaño muestral n tiende al valor $n_0 = PQ/e^2 = f(P)$, expresión que tenemos que maximizar en P. Si igualamos la primera derivada al valor cero tenemos que como $f(P) = P(1-P)/e^2$ entonces $f'(P) = (1-2P)/e^2 = 0 \Rightarrow P = 1/2$. Por otra parte $f''(P) = -2/e^2 < 0$, lo que asegura la presencia de un máximo para la función f en el punto $P = 1/2$. Como $Q = 1-P = 1-1/2 = 1/2$, el valor máximo de n para poblaciones grandes o fracciones de muestreo pequeñas se obtiene para $P = Q = 1/2$. Por lo tanto, para un error prefijado se necesitarán tamaños de muestra más pequeños cuanto más próximo esté P a cero o a uno. Este resultado es muy importante en la práctica, ya que ***cuando se estiman proporciones y no se conoce el valor de la proporción poblacional P ni se tiene una aproximación suya (proporcionada por una encuesta similar, por una encuesta piloto, por la misma encuesta realizada anteriormente o por cualquier otro método), entonces se toma P=1/2***, con lo que estamos situándonos en el caso de máximo tamaño muestral para el error fijado, lo cual siempre es aceptable estadísticamente. La dificultad práctica puede ser que se obtenga un tamaño muestral n demasiado grande para el presupuesto de que se dispone.

Total de clase:

Si sustituimos el valor de S^2 para variables A_i (que sólo toman los valores 0 y 1) en la fórmula del tamaño muestral para el total tendremos para la estimación del total de clase el tamaño:

$$n = \frac{N^2S^2}{e^2 + NS^2} = \frac{N^2\dfrac{N}{N-1}PQ}{e^2 + \dfrac{N}{N-1}PQN} = \frac{N^3PQ}{e^2(N-1) + N^2PQ}$$

También puede estudiarse el ***tamaño de muestra necesario para cometer un error relativo de muestreo*** $e_r = Cv(\hat{\theta})$ dependiendo de si se estima la media, el total, la proporción y el total de clase.

Asimismo, es típico introducir un coeficiente de confianza adicional P_α al error de muestreo a cometer (*límite de tolerancia*). En este caso las fórmulas de los ***tamaños muestrales necesarios para cometer un error absoluto o relativo de muestreo dado en presencia del coeficiente de confianza adicional*** se derivarán de las expresiones $e_\alpha = \lambda_\alpha \sigma(\hat{\theta})$ y $e_{r\alpha} = \lambda_\alpha Cv(\hat{\theta})$. En general $\lambda_\alpha = F^{-1}(1-\alpha/2)$, siendo F la función de distribución de una normal (0,1).

El cuadro siguiente resume las expresiones de los tamaños muestrales.

Tipo de error → *Parámetro* ↓	*Absoluto* e	*Relativo* e_r	*Absoluto y coeficiente de confianza adicional* e_α	*Relativo y confianza* $e_{r\alpha}$
Media	$\frac{NS^2}{Ne^2 + S^2}$	$\frac{NC_{1,x}^2}{Ne_r^2 + C_{1,x}^2}$	$\frac{\lambda_\alpha^2 NS^2}{Ne^2 + \lambda_\alpha^2 S^2}$	$\frac{\lambda_\alpha^2 NC_{1,x}^2}{Ne_{r\alpha}^2 + \lambda_\alpha^2 C_{1,x}^2}$
Total	$\frac{N^2 S^2}{e^2 + NS^2}$	$\frac{NC_{1,x}^2}{Ne_r^2 + C_{1,x}^2}$	$\frac{\lambda_\alpha^2 N^2 S^2}{e^2 + \lambda_\alpha^2 NS^2}$	$\frac{\lambda_\alpha^2 NC_{1,x}^2}{Ne_{r\alpha}^2 + \lambda_\alpha^2 C_{1,x}^2}$
Proporción	$\frac{NPQ}{e^2(N-1) + PQ}$	$\frac{NQ}{P(N-1)e_r^2 + Q}$	$\frac{\lambda_\alpha^2 NPQ}{e^2(N-1) + \lambda_\alpha^2 PQ}$	$\frac{NQ\lambda_\alpha^2}{e_{r\alpha}^2(N-1)P + \lambda_\alpha^2 Q}$
Total de clase	$\frac{N^3 PQ}{e^2(N-1) + N^2 PQ}$	$\frac{NQ}{P(N-1)e_r^2 + Q}$	$\frac{\lambda_\alpha^2 N^3 PQ}{e^2(N-1) + \lambda_\alpha^2 N^2 PQ}$	$\frac{NQ\lambda_\alpha^2}{e_{r\alpha}^2(N-1)P + \lambda_\alpha^2 Q}$

En todas las fórmulas S^2 es la cuasivarianza poblacional y $C_{1,x}^2 = \left(S / \overline{X}\right)^2$. Por otra parte, λ_α es el valor crítico de la normal unitaria al nivel α.

MUESTREO ALEATORIO SIMPLE CON REPOSICIÓN. ESTIMADORES

El muestreo aleatorio simple con reposición es un procedimiento de selección con probabilidades iguales que consiste en obtener la muestra unidad a unidad de forma aleatoria con reposición a la población de las unidades previamente seleccionadas. De esta forma las muestras con elementos repetidos son posibles y cualquier elemento de la población puede estar repetido en la muestra 0, 1, ..., n veces. Supongamos en todo momento que el tamaño de la población es N y el tamaño de la muestra es n. Como la muestra se selecciona con reposición (se reponen a la población las unidades previamente seleccionadas) y con probabilidades iguales, se realiza la selección sucesiva de las unidades para la muestra con probabilidades $P_i = 1/N$ y todas las muestras son equiprobables, ya que:

$$P(u_1, u_2, ..., u_n) = P(u_1)P(u_2) \dots P(u_n) = (1/N)(1/N) \dots (1/N) = 1/(N^n)$$

En cuanto a los estimadores, partimos de que el estimador lineal insesgado general para el caso de muestreo con reposición es el estimador de Hansen y Hurwitz $\hat{\theta}_{HH} = \sum_{i=1}^{n} \frac{Y_i}{nP_i}$ (P_i = probabilidad de seleccionar la unidad u_i de la población para la muestra $=1/N$), que estima insesgadamente la característica poblacional $\theta = \sum_{i=1}^{N} Y_i$. Según los distintos valores de Y_i se tiene:

$$\theta = X = \sum_{i=1}^{N} X_i \Rightarrow Y_i = X_i \Rightarrow \hat{\theta} = \hat{X} = \sum_{i=1}^{n} \frac{X_i}{nP_i} = \sum_{i=1}^{n} \frac{X_i}{\frac{n}{N}} = N \underbrace{\frac{1}{n} \sum_{i=1}^{n} X_i}_{\bar{x}} = N\bar{x}$$

$$\theta = \bar{X} = \sum_{i=1}^{N} \frac{X_i}{N} \Rightarrow Y_i = \frac{X_i}{N} \Rightarrow \hat{\theta} = \hat{\bar{X}} = \sum_{i=1}^{n} \frac{\frac{X_i}{N}}{nP_i} = \sum_{i=1}^{n} \frac{\frac{X_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{x}$$

$$\theta = P = \sum_{i=1}^{N} \frac{A_i}{N} \Rightarrow Y_i = \frac{A_i}{N} \Rightarrow \hat{\theta} = \hat{P} = \sum_{i=1}^{n} \frac{\frac{A_i}{N}}{\frac{n}{N}} = \frac{1}{n} \sum_{i=1}^{n} A_i$$

$$\theta = A = \sum_{i=1}^{N} A_i \Rightarrow Y_i = A_i \Rightarrow \hat{\theta} = \hat{A} = \sum_{i=1}^{n} \frac{A_i}{\frac{n}{N}} = N \frac{1}{n} \sum_{i=1}^{n} A_i = N\hat{P}$$

Da la casualidad de que se obtienen los mismos estimadores insesgados para los parámetros poblacionales que para el caso de muestreo aleatorio simple sin reposición. Por lo tanto, los estimadores de la media y la proporción poblacional son los estimadores por analogía (media y proporción muestral), mientras que los estimadores del total y el total de clase poblacionales son la expansión mediante el tamaño poblacional de la media y proporción muestrales (en este caso, $\hat{X} = N\bar{x} = (N/n)x \Rightarrow$ los factores de elevación son N/n).

VARIANZAS Y SU ESTIMACIÓN CON REPOSICIÓN

Partiendo de la varianza del estimador de Hansen y Hurwitz:

$$V\left(\hat{\theta}_{HH}\right) = \frac{1}{n} \sum_{i=1}^{N} \left(\frac{Y_i}{P_i} - Y \right)^2 P_i$$

y considerando que para el caso particular del muestreo aleatorio simple con reposición se sabe que $P_i = 1/N$, tenemos:

$$V(\hat{X}) = \frac{1}{n} \sum_{i=1}^{N} \left(\frac{X_i}{P_i} - X \right)^2 P_i = \frac{1}{n} \sum_{i=1}^{N} \left(\frac{X_i}{\frac{1}{N}} - X \right)^2 \frac{1}{N} = \frac{N^2}{n} \frac{1}{N} \sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2 = N^2 \frac{\sigma^2}{n}$$

$$V\left(\hat{\bar{X}}\right) = Var\left(\frac{\hat{X}}{N}\right) = \frac{1}{N^2} Var\left(\hat{X}\right) = \frac{1}{N^2} N^2 \frac{\sigma^2}{n} = \frac{\sigma^2}{n}$$

$$V\left(\hat{P}\right) = \frac{\sigma^2}{n} = \frac{\frac{N-1}{N} S^2}{n} = \frac{PQ}{n} \qquad V\left(\hat{A}\right) = N^2 \frac{\sigma^2}{n} = N^2 \frac{PQ}{n}$$

Para estimar las varianzas partimos del estimador de la varianza de Hansen y Hurwitz:

$$\hat{V}(\hat{\theta}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{Y_i}{P_i} - \hat{Y}_{HH} \right)^2$$

y considerando que para el caso particular del muestreo aleatorio simple con reposición se sabe que $P_i = 1/N$, tenemos:

$$\hat{V}\left(\hat{X}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{X_i}{\frac{1}{N}} - \hat{X}\right)^2 = \frac{N^2}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{x}\right)^2 = N^2\frac{\hat{S}^2}{n}$$

$$\hat{V}\left(\hat{\bar{X}}\right) = \hat{V}\left(\frac{\hat{X}}{N}\right) = \frac{1}{N^2}\hat{V}\left(\hat{X}\right) = \frac{1}{N^2}N^2\frac{\hat{S}^2}{n} = \frac{\hat{S}^2}{n}$$

$$\hat{V}\left(\hat{P}\right) = \frac{\hat{S}^2}{n} = \frac{\frac{n}{n-1}\hat{P}\hat{Q}}{n} = \frac{1}{n-1}\hat{P}\hat{Q} \qquad \hat{V}\left(\hat{A}\right) = N^2\frac{\hat{S}^2}{n} = N^2\frac{\frac{n}{n-1}\hat{P}\hat{Q}}{n} = N^2\frac{1}{n-1}\hat{P}\hat{Q}$$

Un resultado interesante que se deduce de las fórmulas anteriores es que *la cuasivarianza muestral definida $\hat{S}^2$ es un estimador insesgado de la varianza poblacional σ^2 en muestreo aleatorio simple con reposición*

TAMAÑO DE LA MUESTRA CON REPOSICIÓN

Igual que en el caso de sin reposición, consideraremos el ***tamaño de muestra necesario para cometer un error de muestreo*** $e = \sigma(\hat{\theta})$ dependiendo de si $\hat{\theta}$ estima la media, el total, la proporción o el total de clase. También se considerará el ***tamaño de muestra necesario para cometer un error relativo de muestreo*** $e_r = Cv(\hat{\theta})$ dependiendo de si se estima la media, el total, la proporción y el total de clase. Asimismo, se tendrá presente la introducción de un coeficiente de confianza adicional P_α al error de muestreo a cometer (*límite de tolerancia*), en cuyo caso las fórmulas de los ***tamaños muestrales necesarios para cometer un error absoluto o relativo de muestreo dado en presencia del coeficiente de confianza adicional*** se derivarán de las expresiones $e_\alpha = \lambda_\alpha\sigma(\hat{\theta})$ y $e_{r\alpha} = \lambda_\alpha Cv(\hat{\theta})$. En general, $\lambda_\alpha = F^{-1}(1-\alpha/2)$, siendo F la función de distribución de una normal (0,1). El cuadro siguiente resume las expresiones de los tamaños muestrales.

Tipo de error → *Parámetro* ↓	*Absoluto* e	*Relativo* e_r	*Absoluto y coeficiente de confianza adicional* e_α	*Relativo y confianza* $e_{r\alpha}$
Media	$\frac{\sigma^2}{e^2}$	$\frac{C_x^2}{e_r^2}$	$\frac{\lambda_\alpha^2\sigma^2}{e^2}$	$\frac{\lambda_\alpha^2 C_x^2}{e_{r\alpha}^2}$
Total	$\frac{N^2\sigma^2}{e^2}$	$\frac{C_x^2}{e_r^2}$	$\frac{\lambda_\alpha^2 N^2\sigma^2}{e^2}$	$\frac{\lambda_\alpha^2 C_x^2}{e_{r\alpha}^2}$
Proporción	$\frac{PQ}{e^2}$	$\frac{Q}{Pe_r^2}$	$\frac{\lambda_\alpha^2 PQ}{e^2}$	$\frac{\lambda_\alpha^2 Q}{Pe_{r\alpha}^2}$
Total de clase	$\frac{N^2 PQ}{e^2}$	$\frac{Q}{Pe_r^2}$	$\frac{\lambda_\alpha^2 N^2 PQ}{e^2}$	$\frac{\lambda_\alpha^2 Q}{Pe_{r\alpha}^2}$

En todas las fórmulas σ^2 es la varianza poblacional y $C_x^2 = \left(\sigma/\bar{X}\right)^2$. Por otra parte, λ_α es el valor crítico de la normal unitaria al nivel α.

COMPARACIÓN ENTRE MUESTREO ALEATORIO SIN Y CON REPOSICIÓN

Se pueden realizar las comparaciones a través error de muestreo o a través del tamaño muestral necesario para cometer un error de muestreo dado. Desde el primer enfoque será más preciso aquel método de selección cuyo error de muestreo sea menor, es decir, el que tenga menor varianza de los estimadores. Tenemos:

$$\left.\begin{aligned} V_{SR}\left(\hat{\bar{X}}\right) &= (1-f)\frac{S^2}{n} = (1-\frac{n}{N})\frac{\frac{N}{N-1}\sigma^2}{n} = \frac{N-n}{N-1}\frac{\sigma^2}{n} \\ V_{CR}\left(\hat{\bar{X}}\right) &= \frac{\sigma^2}{n} \Rightarrow n = \frac{\sigma^2}{e^2} \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} &\frac{V_{SR}\left(\hat{\bar{X}}\right)}{V_{CR}\left(\hat{\bar{X}}\right)} = \frac{N-n}{N-1} < 1 \\ &\Rightarrow V_{SR}\left(\hat{\bar{X}}\right) < V_{CR}\left(\hat{\bar{X}}\right) \end{aligned}\right.$$

Para el resto de los estimadores todo sería equivalente, luego la varianza siempre es menor en el caso del muestreo sin reposición, lo que nos indica que ***el muestreo sin reposición es en general más preciso que el muestreo con reposición.***

Desde el punto de vista del tamaño muestral, será mejor aquel método de selección en el que se necesite menor tamaño muestral para cometer un error de muestreo dado. En este capítulo hemos visto que para muestreo sin reposición el valor de *n* era:

$$n_{SR} = \frac{n_0}{1 + n_0/N}$$

tanto en el caso de estimaciones de medias y proporciones para un error de muestreo dado como en el caso de estimaciones de medias, totales, proporciones y totales de clase para un error relativo de muestreo dado con o sin coeficiente de confianza. En los mismos casos, para muestreo con reposición se observa que el tamaño muestral resulta ser $n_{CR} = n_0$. Por lo tanto, tenemos:

$$n_{SR} = \frac{n_0}{1 + n_0/N} = \frac{n_{CR}}{1 + n_{CR}/N} < n_{CR} \Rightarrow n_{SR} < n_{CR}$$

En el caso de estimación sin reposición de totales y totales de clase para un error de muestreo dado con o sin coeficiente de confianza se vio que:

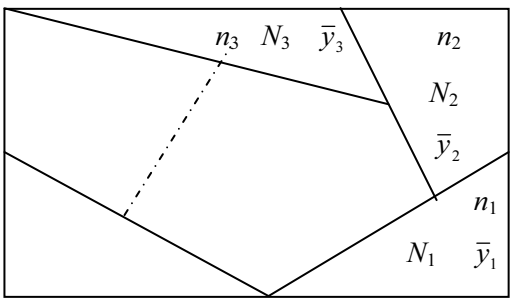
$$n_{SR} = \frac{N^2 n_1}{1 + N n_1} \cong \frac{n_{CR}}{1 + n_{CR}/N} < n_{CR} \Rightarrow n_{SR} < n_{CR}$$

En los mismos casos para muestreo con reposición se observa que el tamaño muestral resulta ser $n_{CR} = N^2 n_1$.

Por lo tanto, en todas las situaciones, ***en el caso de muestreo sin reposición se necesita menos tamaño de muestra para cometer el mismo error que en el caso del muestreo con reposición***, con lo que el muestreo sin reposición es más eficiente que el muestreo con reposición.

SUBPOBLACIONES

La escasa disponibilidad de marcos que listen específicamente los elementos de la población que interesa estudiar, sobre todo cuando utilizamos unidades poblacionales muy elementales (marco muy fino), nos lleva a considerar la teoría de subpoblaciones o dominios. Normalmente se dispone de marcos menos finos cuyas unidades contienen a las unidades elementales en estudio. Por ejemplo, podemos desear estudiar una muestra de los hogares que tienen niños, pero el mejor marco disponible puede ser una lista de todos los hogares en la ciudad (sin poder desagregar hasta los hogares que tienen niños). Utilizaremos entonces el marco amplio de todos los hogares y consideraremos la subpoblación de los hogares que tienen niños para intentar estimar los parámetros de dicha subpoblación a través de los métodos para subpoblaciones. Supongamos que dividimos una población de tamaño N en subpoblaciones o dominios. Consideremos que el j-ésimo dominio contiene N_j unidades, y que n_j es el número de unidades, en una muestra aleatoria simple de tamaño n, que pertenecen al dominio j.



N = Tamaño de la población

n = tamaño de la muestra

Sea Y_{jk} (k = 1, 2, ..., n_j y $\sum n_j = n$) son los valores de la variable en estudio medida sobre los elementos de la muestra que pertenecen al dominio j-ésimo. Un ***estimador insesgado de la media en la subpoblación o dominio j*** será el siguiente:

$$\hat{\bar{Y}}_j = \bar{y}_j = \sum_{k=1}^{n_j} \frac{Y_{jk}}{n_j}$$

cuya varianza puede expresarse como:

$$V(\bar{y}_j) = (1 - \frac{n_j}{N_j})\frac{S_j^2}{n_j} \text{ siendo } S_j^2 = \frac{1}{N_j - 1}\sum_{k=1}^{N_j}\left(Y_{jk} - \bar{Y}_j\right)^2 \text{ donde } \bar{Y}_j = \sum_{k=1}^{N_j}\frac{Y_{jk}}{N_j}$$

y pudiendo expresarse la estimación de su varianza como:

$$\hat{V}(\bar{y}_j) = (1 - \frac{n_j}{N_j})\frac{\hat{S}_j^2}{n_j} \text{ siendo } \hat{S}_j^2 = \frac{1}{n_j - 1}\sum_{k=1}^{n_j}\left(Y_{jk} - \bar{y}_j\right)^2 \text{ dónde } \bar{y}_j = \sum_{k=1}^{n_j}\frac{Y_{jk}}{n_j}$$

Si no se conoce el valor de N_j, se sustituye n_j/N_j por n/N y se tiene:

$$V(\bar{y}_j) = (1 - \frac{n}{N})\frac{S_j^2}{n_j} \qquad \hat{V}(\bar{y}_j) = (1 - \frac{n}{N})\frac{\hat{S}_j^2}{n_j}$$

En el caso del muestreo con reposición tenemos:

$$V(\bar{y}_j) = \frac{\sigma_j^2}{n_j} \text{ siendo } \sigma_j^2 = \frac{1}{N_j}\sum_{k=1}^{N_j}\left(Y_{jk} - \bar{Y}_j\right)^2 \text{ y } \hat{V}(\bar{y}_j) = \frac{\hat{S}_j^2}{n_j}$$

Un ***estimador insesgado del total en la subpoblación o dominio j en caso de conocer N_j*** será el siguiente:

$$\hat{Y}_j = N_j\bar{y}_j = N_j\sum_{k=1}^{n_j}\frac{Y_{jk}}{n_j}$$

cuya varianza y estimación de varianza son, respectivamente:

$$V(\hat{Y}_j) = N_j^2 V(\bar{y}_j) = N_j^2(1 - \frac{n_j}{N_j})\frac{S_j^2}{n_j} \quad \text{y} \quad \hat{V}(\hat{Y}_j) = N_j^2(1 - \frac{n_j}{N_j})\frac{\hat{S}_j^2}{n_j}$$

En el muestreo con reposición tendremos:

$$V(\hat{Y}_j) = N_j^2 V(\bar{y}_j) = N_j^2\frac{\sigma_j^2}{n_j} \quad \text{y} \quad \hat{V}(\hat{Y}_j) = N_j^2\frac{\hat{S}_j^2}{n_j}$$

Un ***estimador insesgado del total en la subpoblación o dominio j en caso de no conocer N_j*** será el siguiente:

$$\hat{Y}_j = N_j\sum_{k=1}^{n_j}\frac{Y_{jk}}{n_j} = \sum_{k=1}^{n_j}\frac{N_j}{n_j}Y_{jk} \underbrace{=}_{\substack{\text{Se aplica} \\ \frac{N_j}{n_j}\to\frac{N}{n}}} \frac{N}{n}\sum_{k=1}^{n_j}Y_{jk} = \frac{N}{n}\underbrace{y_j}_{\substack{\text{Total} \\ \text{muestral} \\ \text{en dominio} \\ j-\text{ésimo}}}$$

cuya varianza y estimación de varianza son, respectivamente:

$$V(\hat{Y}_j) = N^2(1 - \frac{n}{N})\frac{S'^2}{n} \quad \text{y} \quad \hat{V}(\hat{Y}_j) = N^2(1 - \frac{n}{N})\frac{\hat{S}'^2}{n}$$

siendo $S'^2 = \frac{1}{N-1}\left(\sum_{Dominio\,j} {Y_{jk}}^2 - \frac{Y_j^2}{N}\right)$, $\hat{S}'^2 = \frac{1}{n-1}\left(\sum_{k=1}^{n_j} {Y_{jk}}^2 - \frac{y_j^2}{n}\right)$ e $y_j = \sum_{k=1}^{n_j}Y_{jk}$

En el muestreo con reposición tendremos:

$$V(\hat{Y}_j) = N^2\frac{\sigma'^2}{n} \quad \text{y} \quad \hat{V}(\hat{Y}_j) = N^2\frac{\hat{S}'^2}{n} \quad \text{con} \quad \sigma'^2 = \frac{1}{N}\left(\sum_{Dominio\ j} {Y_{jk}}^2 - \frac{Y_j^2}{N}\right)$$

PROBLEMAS RESUELTOS

3.1. Un auditor muestrea aleatoriamente con reposición 20 cuentas impagadas de una empresa y verifica en 12 de ellas la cantidad adeudada y si los documentos respectivos cumplen (1) o no cumplen (0) con los procedimientos establecidos. Se tienen la siguiente estructura poblacional:

Cuenta	Cantidad	Concordancia	Cuenta	Cantidad	Concordancia
1	278	1	11	188	0
2	192	1	12	212	0
3	310	1	13	92	1
4	94	0	14	56	1
5	86	1	15	142	1
6	335	1	16	37	1
7	310	0	17	186	0
8	290	1	18	221	1
9	221	1	19	229	0
10	168	1	20	305	1

Basándose en las 12 cuentas verificadas, estimar la proporción de cuentas cuyos documentos concuerdan, así como el importe medio adeudado, y cuantificar el error cometido.

Comenzamos introduciendo los datos en una hoja de cálculo de Excel. A continuación, para elegir la muestra, en el menú *Herramientas* de Excel elegimos *Análisis de datos*, seleccionamos *Muestra* y rellenamos la pantalla de entrada como se indica en la Figura 3-1. Al pulsar *Aceptar* se obtiene la MUESTRA de tamaño 12 de la Figura 3-2. Mediante las fórmulas de la Figura 3-2 se obtienen los resultados de la Figura 3-3.

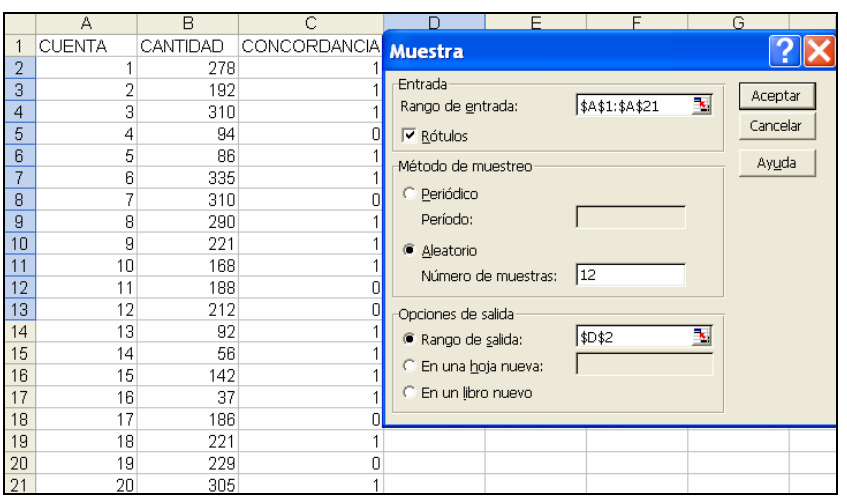


Figura 3-1

	A	B	C	D	E	F
1	CUENTA	CANTIDAD	CONCORDANCIA	**MUESTRA**	X	A
2	1	278	1	**17**	186	0
3	2	192	1	**18**	221	1
4	3	310	1	**4**	94	0
5	4	94	0	**5**	86	1
6	5	86	1	**1**	278	1
7	6	335	1	**6**	335	1
8	7	310	0	**2**	192	1
9	8	290	1	**7**	310	0
10	9	221	1	**14**	56	1
11	10	168	1	**20**	305	1
12	11	188	0	**3**	310	1
13	12	212	0	**15**	142	1
14	13	92	1	**Estimaciones:**	**=PROMEDIO(X)**	**=PROMEDIO(A)**
15	14	56	1	**Errores absolutos:**	**=VARP(CANTIDAD)/12**	**=VARP(CONCORDANCIA)/12**
16	15	142	1	**Errores relativos:**	**=100*RAIZ(E15)/E14**	**=100*RAIZ(F15)/F14**
17	16	37	1			
18	17	186	0			
19	18	221	1			
20	19	229	0			
21	20	305	1			

Figura 3-2

	A	B	C	D	E	F
1	CUENTA	CANTIDAD	CONCORDANCIA	MUESTRA	X	A
2	1	278	1	17	186	0
3	2	192	1	18	221	1
4	3	310	1	4	94	0
5	4	94	0	5	86	1
6	5	86	1	1	278	1
7	6	335	1	6	335	1
8	7	310	0	2	192	1
9	8	290	1	7	310	0
10	9	221	1	14	56	1
11	10	168	1	20	305	1
12	11	188	0	3	310	1
13	12	212	0	15	142	1
14	13	92	1	**Estimaciones:**	**209,5833333**	**0,75**
15	14	56	1	**Errores absoluto**	**655,745**	**0,0175**
16	15	142	1	**Errores relativos**	**12,21829905**	**17,6383421**
17	16	37	1			
18	17	186	0			
19	18	221	1			
20	19	229	0			
21	20	305	1			

Figura 3-3

Hemos obtenido que el importe medio adeudado se estima en:

$$\overline{X} = \frac{1}{12}\sum_{i=1}^{12} X_i = 209{,}583$$

con un error absoluto de:

$$V\left(\hat{\overline{X}}\right) = \frac{\sigma^2}{n} = 655{,}745$$

La proporción de cuentas cuyos documentos concuerdan con los procedimientos establecidos se estima mediante:

$$\hat{P} = \frac{1}{12}\sum_{i=1}^{12} A_i = 0{,}75$$

El error absoluto de esta estimación es:

$$V\left(\hat{P}\right) = \frac{PQ}{n} = \frac{(14/20)(6/20)}{12} = \frac{84}{4800} = 0{,}0175$$

En términos relativos estos errores son más fácilmente interpretables, y se cuantifican en el 12,2% y el 17,6%, respectivamente (a través de los coeficientes de variación de los estimadores). Tenemos:

$$Cv\left(\hat{\bar{X}}\right)=\frac{\sqrt{655{,}745}}{209{,}583}100=12{,}218\% \qquad Cv\left(\hat{P}\right)=\frac{\sqrt{0{,}0175}}{0{,}75}100=17{,}638\%$$

3.2. El gerente de un taller de maquinaria desea estimar el tiempo promedio que necesita un operador para terminar una tarea sencilla. El taller tiene 98 operadores y se selecciona una muestra de 8 sin reposición a los que se les toma el tiempo, Se obtienen los siguientes resultados:

4,2 5,1 7,9 3,8 5,3 4,6 5,1 4,1

Estimar el tiempo promedio y el tiempo total para terminar la tarea entre todos los operadores estableciendo límites al 95% para los errores de estimación.

Comenzamos introduciendo los datos como la variable *T* en una hoja de cálculo de Excel. A continuación, para calcular los estadísticos necesarios, en el menú *Herramientas* de Excel elegimos *Análisis de datos*, seleccionamos *Estadística descriptiva* y rellenamos la pantalla de entrada como se indica en la Figura 3-4. Al pulsar *Aceptar* se obtienen los estadísticos maestrales de la Figura 3-5. Por último, se calculan los estimadores y sus errores según las fórmulas de la Figura 3-6 que nos llevan a los resultados de la Figura 3-7.

Se observa que el tiempo medio por operario para terminar la tarea es $\hat{\bar{T}}=\frac{1}{12}\sum_{i=1}^{12}T_i=$ 5,0125 minutos con un error de muestreo de $\hat{V}(\hat{\bar{T}})=(1-f)\frac{\hat{S}^2}{n}=$ 0,189 y un error relativo dado por $Cv(\hat{\bar{T}})=\frac{\hat{\sigma}(\hat{\bar{T}})}{\hat{\bar{T}}}100=\frac{\sqrt{0{,}189}}{5{,}0125}100=$ 9,69%. El tiempo total para terminar la tarea se estima en $\hat{T}=N\hat{\bar{T}}=89\cdot 5{,}0125=$ 491,225 minutos con un error de muestreo estimado por $\hat{V}(\hat{T})=N^2\hat{V}(\hat{\bar{T}})=89^2\cdot 0{,}189=$ 1822,07, siendo el error relativo el mismo que el del estimador del tiempo medio, es decir, 9,69%. El coeficiente de curtosis = 4,24 no está en el intervalo [−2,2] luego no podemos suponer normalidad, con lo que intervalo de confianza al 95% para la media de anchura 1,07475886 no es válido.

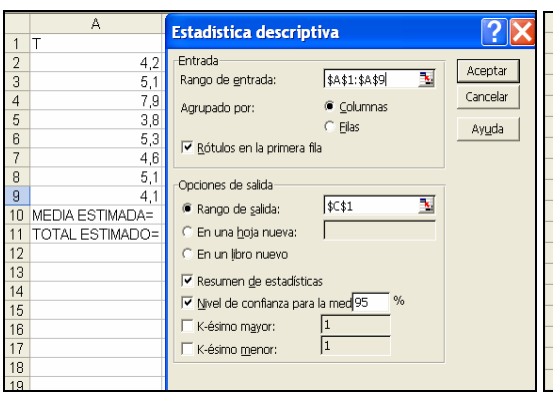


Figura 3-4

	A	B	C	D
1	T		*T*	
2	4,2			
3	5,1		Media	5,0125
4	7,9		Error típico	0,45451603
5	3,8		Mediana	4,85
6	5,3		Moda	5,1
7	4,6		Desviación estándar	1,28556547
8	5,1		Varianza de la muestra	1,65267857
9	4,1		Curtosis	4,24373466
10	MEDIA ESTIMADA=		Coeficiente de asimetría	1,87830493
11	TOTAL ESTIMADO=		Rango	4,1
12			Mínimo	3,8
13			Máximo	7,9
14			Suma	40,1
15			Cuenta	8
16			Nivel de confianza(95,0%)	1,07475886
17				

Figura 3-5

	A	B	C	D
1	T		*T1*	
2	4,2			
3	5,1		Media	5,0125
4	7,9		Error típico	0,454516029891763
5	3,8		Mediana	4,85
6	5,3		Moda	5,1
7	4,6		Desviación estándar	1,28556546757781
8	5,1		Varianza de la muestra	1,65267857142856
9	4,1		Curtosis	4,24373466005058
10	**MEDIA ESTIMADA=**	**=PROMEDIO(T)**	Coeficiente de asimetría	1,87830492924543
11	**TOTAL ESTIMADO=**	**=98*PROMEDIO(T)**	Rango	4,1
12	**VAR(MEDIA)=**	**=(1-8/98)*($D$8)/8**	Mínimo	3,8
13	**VAR(TOTAL)=**	**=98^2*$B$12**	Máximo	7,9
14	**ER(MEDIA)=**	**=100*RAIZ(B12)/B10**	Suma	40,1
15	**ER(TOTAL)=**	**=100*RAIZ(B13)/B11**	Cuenta	8
16	**CONFIANZA(MEDIA)=**	**=2*RAIZ(B12/0,05)**	Nivel de confianza(95,0%)	1,07475885815483
17	**CONFIANZA(TOTAL)=**	**=2*RAIZ(B13/0,05)**		

Figura 3-6

	A	B	C	D
1	T		*T1*	
2	4,2			
3	5,1		Media	5,0125
4	7,9		Error típico	0,45451603
5	3,8		Mediana	4,85
6	5,3		Moda	5,1
7	4,6		Desviación estándar	1,28556547
8	5,1		Varianza de la muestra	1,65267857
9	4,1		Curtosis	4,24373466
10	**MEDIA ESTIMADA=**	**5,0125**	Coeficiente de asimetría	1,87830493
11	**TOTAL ESTIMADO=**	**491,225**	Rango	4,1
12	**VAR(MEDIA)=**	**0,189720754**	Mínimo	3,8
13	**VAR(TOTAL)=**	**1822,078125**	Máximo	7,9
14	**ER(MEDIA)=**	**8,689665035**	Suma	40,1
15	**ER(TOTAL)=**	**8,689665035**	Cuenta	8
16	**CONFIANZA(MEDIA)=**	**3,895851685**	Nivel de confianza(95,0%)	1,07475886
17	**CONFIANZA(TOTAL)=**	**381,7934651**		

Figura 3-7

Al no existir normalidad utilizamos como intervalos de confianza:

$$\left[\hat{\theta}-\frac{\sigma\left(\hat{\theta}\right)}{\sqrt{\alpha}},\hat{\theta}+\frac{\sigma\left(\hat{\theta}\right)}{\sqrt{\alpha}}\right]$$

cuya anchura es $2\dfrac{\sigma\left(\hat{\theta}\right)}{\sqrt{\alpha}}$. Esta anchura (3,895 para el estimador de la media y 381,79 para el estimador del total) suele considerarse como un límite para el error de estimación. Se observa que estas anchuras son mayores que con normalidad, ya que en este caso las estimaciones son menos precisas (errores mayores).

3.3. En una región con $N = 1000$ viviendas determinar el tamaño de muestra necesario para que, con un grado de confianza del 95%, la estimación de la proporción de viviendas sin agua corriente no difiera en más del 0,1 del valor verdadero. Comentar los resultados para muestreo sin reposición y con reposición.

$$P(|\hat{P}-P|\leq 0{,}10)=0{,}95 \Leftrightarrow P(-0{,}10\leq \hat{P}-P\leq 0{,}10)=0{,}95 \Leftrightarrow$$

$$P\left(\frac{-0{,}10}{\sigma(\hat{P})}\leq \frac{\hat{P}-P}{\sigma(\hat{P})}\leq \frac{0{,}10}{\sigma(\hat{P})}\right)=0{,}95 \Leftrightarrow P\left(\frac{-0{,}10}{\sigma(\hat{P})}\leq N(0{,}1)\leq \frac{0{,}10}{\sigma(\hat{P})}\right)=0{,}95$$

De lo anterior se deduce que:

$$\frac{0{,}10}{\sigma(\hat{P})}=\lambda_\alpha=1{,}96 \Rightarrow \sigma(\hat{P})=\frac{0{,}10}{1{,}96}=0{,}051$$

Luego el problema se traduce en calcular el tamaño de muestra necesario para cometer un error de muestreo de 0,051 al estimar la proporción de viviendas sin agua corriente. *Como no tenemos información acerca de la proporción poblacional P de viviendas sin agua corriente, nos colocamos en la situación más desfavorable, es decir, P = Q* = 1/2. Tendremos:

$$n=\frac{NP(1-P)}{P(1-P)+(N-1)e^2}=\frac{1000\cdot 0{,}5\cdot 0{,}5}{0{,}5\cdot 0{,}5+999.0{,}051^2}=91 \text{ viviendas}$$

Para el caso de muestreo con reposición tendremos:

$$n=\frac{P(1-P)}{e^2}=\frac{0{,}5\cdot 0{,}5}{0{,}051^2}=96 \text{ viviendas}$$

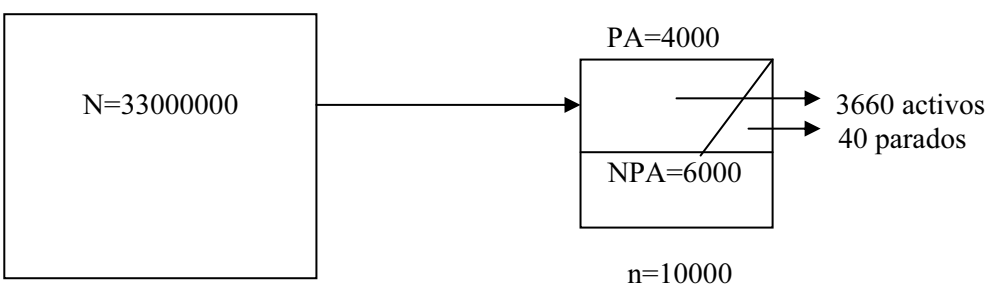
Se observa que el tamaño de muestra necesario para cometer el mismo error de muestreo al estimar igual parámetro es superior en el caso de muestreo con reposición.

3.4. De una población con 33 millones de habitantes se ha obtenido una muestra de 10.000. En ella, 4.000 se han clasificado como población activa, y de éstos, 40 se encuentran en situación de desempleo. Se pide:

1) Estimar el porcentaje de población activa. Estimar también el número de personas activas que se encuentran en situación de desempleo. Calcular los errores absoluto y relativo de muestreo en ambas estimaciones así como intervalos de confianza con un riesgo del 3 por mil.

2) ¿Cuántas personas de todas las edades sería necesario incluir en una muestra para estimar la tasa de actividad en España con un error absoluto E = 0,02 y una probabilidad del 95%? Del último censo se sabe que en el país hay un 39% de activos. Contestar a la misma pregunta para cometer un error relativo del 5%.

Realizamos el siguiente esquema de apoyo (PA significa población activa y NPA significa el complentario):



El porcentaje estimado de población activa será:

$$\hat{P} = \frac{4000}{10000} = 0{,}4 \quad (40\%)$$

El error de muestreo será:

$$\hat{\sigma}(\hat{P}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{P}(1-\hat{P})}{n-1}} = \sqrt{\left(1 - \frac{10000}{33000000}\right)\frac{0{,}4(1-0{,}4)}{10000-1}} = 0{,}00489$$

El error relativo de muestreo será la estimación del coeficiente de variación de $\hat{P}$, que se calcula de la siguiente forma:

$$\hat{C}v(\hat{P}) = \frac{\hat{\sigma}(\hat{P})}{\hat{P}} = \frac{0{,}00489}{0{,}4} = 0{,}012225 \quad (1{,}2225\%)$$

Para hallar el intervalo de confianza para la proporción con $\alpha = 0{,}003$, utilizamos $\lambda_\alpha = F^{-1}_{N(0,1)}(1-\alpha/2) = F^{-1}_{N(0,1)}(1-0{,}003/2) = F^{-1}_{N(0,1)}(0{,}9985) = 2{,}997$. El intervalo será:

$$[\hat{P} - \lambda_\alpha \sigma(\hat{P}), \hat{P} + \lambda_\alpha \sigma(\hat{P})] = [0{,}4 - 2{,}997 \cdot 0{,}00489, 0{,}4 + 2{,}997 \cdot 0{,}00489] = (0.3853, 0.4146)$$

Se podría interpretar el intervalo de confianza diciendo que el porcentaje de la población activa está comprendido entre el 38,53% y el 41,46% con una probabilidad del 997 por mil, es decir, prácticamente la certeza.

El total estimado de personas activas que se encuentran en situación de desempleo será:

$$\hat{A} = 33000000\underbrace{\left(\frac{40}{10000}\right)}_{\hat{P}} = 132000$$

El error de muestreo será:

$$\hat{\sigma}(\hat{A}) = \sqrt{N^2\left(1 - \frac{n}{N}\right)\frac{\hat{P}(1-\hat{P})}{n-1}} = 33000000\sqrt{\left(1 - \frac{10000}{33000000}\right)\frac{0{,}004(1-0{,}004)}{10000-1}} = 20827$$

El error relativo de muestreo será la estimación del coeficiente de variación de $\hat{A}$, que se calcula de la siguiente forma:

$$\hat{C}v(\hat{A}) = \frac{\hat{\sigma}(\hat{A})}{\hat{A}} = \frac{20827}{132000} = 0{,}157 \quad (15{,}7\%)$$

Para hallar el intervalo de confianza para el total con $\alpha = 0{,}003$, utilizamos el valor $\lambda_\alpha = F^{-1}_{N(0,1)}(1-\alpha/2) = F^{-1}_{N(0,1)}(1-0{,}003/2) = F^{-1}_{N(0,1)}(0{,}9985) = 2{,}997$. El intervalo será:

$$[\hat{A} - \lambda_\alpha \sigma(\hat{A}), \hat{A} + \lambda_\alpha \sigma(\hat{A})] = [132000 - 2{,}997 \cdot 20827, 132000 + 2{,}997 \cdot 20827] = (69581, 194419)$$

El tamaño de muestra necesario para estimar la tasa de actividad en España con un error de muestreo $e_\alpha = 0{,}02$ y un coeficiente de confianza del 95% será:

$$n = \frac{\lambda_\alpha^2 NPQ}{(N-1)e_\alpha^2 + \lambda_\alpha^2 PQ} = \frac{1{,}96^2 \cdot 33000000 \cdot 0{,}39 \cdot (1-0{,}39)}{(33000000-1) \cdot 0{,}02^2 + 1{,}96^2 \cdot 0{,}39 \cdot (1-0{,}39)} = 2379$$

El tamaño de muestra necesario para estimar la tasa de actividad en España con un error relativo de muestreo $e_{r\,\alpha}$=0,05 y un coeficiente de confianza del 95% será:

$$n = \frac{\lambda_{r\alpha}^2 NQ}{(N-1)Pe_{r\alpha}^2 + \lambda_{r\alpha}^2 Q} = \frac{1{,}96^2 \cdot 33000000 \cdot (1-0{,}39)}{(33000000-1)\cdot 0{,}39 \cdot 0{,}02^2 + 1{,}96^2 \cdot (1-0{,}39)} = 2379$$

3.5. Mediante muestreo irrestricto aleatorio se trata de estimar la proporción y el total de aciertos obtenidos en un juego ilegal en el que se realizan un total de 6000 apuestas. En un ensayo previo se han obtenido 1/3 de fallos en las apuestas. Se pide:

1) Hallar el número de apuestas necesario para que el error de muestreo sea de una décima al estimar la proporción de aciertos en las apuestas del juego ilegal. Hallar también el número de apuestas necesario para que el error relativo de muestreo sea del 20% en la misma estimación.

2) Hallar el número de apuestas necesario para que el error de muestreo sea de 600 unidades al estimar el total de aciertos en las apuestas con un coeficiente de confianza del 99,7% y suponiendo muestreo aleatorio simple con reposición. Hallar dicho tamaño en las condiciones anteriores pero para un error relativo de muestreo del 10%.

Tenemos como datos N = 6000 y P = 2/3. El tamaño de muestra necesario para estimar la proporción de aciertos en las apuestas con un error de muestreo e = 0,1 será:

$$n = \frac{NPQ}{(N-1)e^2 + PQ} = \frac{6000 \cdot 0{,}6666 \cdot (1-0{,}6666)}{(6000-1)\cdot 0{,}1^2 + 0{,}6666 \cdot (1-0{,}6666)} = 22{,}14$$

Será necesario utilizar un tamaño de muestra de 23 apuestas.

El tamaño de muestra necesario para estimar la proporción de aciertos con un error relativo de muestreo e_r = 0,2 será:

$$n = \frac{NQ}{(N-1)Pe_r^2 + Q} = \frac{6000 \cdot (1-0{,}6666)}{(6000-1)\cdot 0{,}6666 \cdot 0{,}2^2 + (1-0{,}6666)} = 12{,}47$$

Será necesario utilizar un tamaño de muestra de 13 apuestas.

Para hallar el tamaño de muestra necesario para estimar el total de aciertos con α = 0,003, se usa $\lambda_\alpha = F^{-1}_{N(0,1)}\,(1-\alpha/2) = F^{-1}_{N(0,1)}\,(1-0{,}003/2) = F^{-1}_{N(0,1)}\,(0{,}9985) = 2{,}997$. Dicho tamaño en muestreo con reposición para un error de muestreo e_α= 600 se calcula de la siguiente forma:

$$n = \frac{\lambda_\alpha^2 PQN^2}{e_\alpha^2} = \frac{2{,}997^2 \cdot 0{,}6666(1-0{,}6666)6000^2}{600^2} = 199{,}6 \quad (200 \text{ apuestas})$$

El tamaño de muestra en muestreo con reposición para un error relativo de muestreo $e_{r\alpha}$= 0,1 con α = 0,003 se calcula de la siguiente forma:

$$n = \frac{\lambda_\alpha^2 Q}{e_\alpha^2 P} = \frac{2{,}997^2 \cdot (1-0{,}6666)}{0{,}1^2 \cdot 0{,}6666} = 449{,}1 \quad (450 \text{ apuestas})$$

3.6. De una población de 100 opositores que se presentan a un examen se ha extraído una muestra irrestricta aleatoria de tamaño $n = 8$, siendo sus edades (variable X) las siguientes: {25, 32, 28, 35, 26, 34, 30, 28}. Basándose en esta muestra, estimar la edad media y la suma de las edades de los opositores así como sus errores absoluto y relativo de muestreo. Determinar también:

1) Basándose en la muestra anterior, ¿qué tamaño de muestra sería necesario para que el error de muestreo sea 2 al estimar la edad media y 50 al estimar la suma de las edades? ¿Y para que el error relativo sea del 6%? Contestar a las mismas preguntas con un coeficiente de confianza del 95%.

2) A partir de la muestra anterior, estimar la proporción de edades pares en la población y el total de la clase de las edades pares estimando los errores absoluto y relativo de muestreo. ¿Qué tamaño de muestra sería necesario para que el error relativo de muestreo fuese del 6% al 95% de confianza al estimar la proporción?

3) Hallar el tamaño de muestra del apartado anterior suponiendo muestreo con reposición. Comentar los resultados.

Se observa que la media muestral es 29,75, la cuasivarianza muestral es 13,3571 y la cuasidesviación típica muestral es 3,65474. También se obtienen buenos valores para los coeficientes de asimetría (0,28) y curtosis (–0,79), que al estar comprendidos entre –2 y 2 permiten suponer normalidad.

Las estimaciones de la edad media y la suma de edades y sus errores absoluto y relativo son:

$$\hat{\bar{X}} = \bar{x} = 29{,}75 \qquad e = \hat{\sigma}(\bar{x}) = \sqrt{(1-f)\frac{\hat{S}^2}{n}} = \sqrt{\left(1-\frac{8}{100}\right)\frac{13.3571}{8}} = 1{,}536$$

$$e_r = Cv(\bar{x}) = \frac{\hat{\sigma}(\bar{x})}{\bar{x}} = \frac{1{,}536}{29{,}75} = 0{,}051 \quad (5{,}1\%)$$

$$\hat{X} = N \cdot \bar{x} = 100 \cdot 29{,}75 = 2975 \qquad e = \hat{\sigma}(\hat{X}) = N \cdot \hat{\sigma}(\bar{x}) = 100 \cdot 1{,}536 = 153{,}6$$

$$e_r = Cv(\hat{X}) = \frac{\hat{\sigma}(\hat{X})}{\hat{X}} = \frac{153{,}6}{2975} = 0{,}051 \quad (5{,}1\%)$$

Evidentemente, los errores relativos de las estimaciones de media y total coinciden.

Para hallar el tamaño de muestra necesario para estimar la edad media (media) con un error de muestreo e igual a 50, consideramos la muestra anterior como una muestra piloto que nos proporciona una estimación del valor de la cuasivarianza. Se aplica la fórmula:

$$n = \frac{NS^2}{S^2 + Ne^2} = \frac{100 \cdot 13{,}3571}{13{,}3571 + 100.2^2} = 3{,}23$$

con lo que se tomará como tamaño de muestra necesario $n = 4$.

Para hallar el tamaño de muestra necesario para estimar la suma de edades (total) con un error de muestreo e igual a 50, se aplica la fórmula:

$$n = \frac{N^2 S^2}{NS^2 + e^2} = \frac{100^2 \cdot 13{,}3571}{100 \cdot 13{,}3571 + 50^2} = 34{,}82$$

con lo que se tomará como tamaño de muestra necesario $n = 35$.

Si introducimos un coeficiente de confianza del 95%, los tamaños de muestra necesarios para cometer el mismo error de muestreo $e_\alpha = 2$ al estimar la media y $e_\alpha = 50$ para el total lógicamente serán algo superiores a los calculados anteriormente. Tenemos:

$$\text{Media} \rightarrow n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{12{,}82}{1 + \frac{12{,}82}{100}} = 11{,}36 \text{ con } n_0 = \frac{\lambda_\alpha^2 S^2}{e_\alpha^2} = \frac{1{,}96^2 \cdot 13{,}3571}{2^2} = 12{,}82$$

$$\text{Total} \rightarrow n = \frac{N^2 n_1}{1 + N n_1} = \frac{100^2 12{,}82}{1 + 100 \cdot 12{,}82} = 99{,}92 \text{ con } n_0 = \frac{\lambda_\alpha^2 S^2}{e_\alpha^2} = \frac{1{,}96^2 \cdot 13{,}3571}{2^2} = 12{,}82$$

Para el caso de un error relativo de muestreo igual a $e_r = 0{,}06$ el tamaño de muestra necesario es el mismo para la estimación del total y de la media. Tendremos:

$$n = \frac{C_{1,x}^2}{e_r^2 + \frac{C_{1,x}^2}{N}} = \frac{0{,}015}{0{,}06^2 + \frac{0{,}015}{100}} = 4 \text{ con } C_{1,x}^2 = \frac{S^2}{\overline{X}^2} = \frac{13{,}3571}{29{,}75^2} = 0{,}015$$

Para el caso de un error relativo de muestreo igual a $e_{r\alpha} = 0{,}06$ con un coeficiente de confianza del 95%, el tamaño de muestra necesario es el mismo para la estimación del total y de la media, y lógicamente será mayor que cuando no existe el coeficiente de confianza. Tendremos:

$$n = \frac{\lambda_\alpha^2 C_{1,x}^2}{e_{r\alpha}^2 + \lambda_\alpha^2 \frac{C_{1,x}^2}{N}} = \frac{1{,}96^2 \cdot 0{,}015}{0{,}06^2 + 1{,}96^2 \cdot \frac{0{,}015}{100}} = 61{,}54$$

con lo que se tomará como tamaño de muestra necesario $n = 65$ que, evidentemente, es superior al tamaño de muestra necesario sin coeficiente de confianza.

A continuación consideramos la muestra asociada a la inicial, cuyos valores son cero para edades impares y uno para edades pares, es decir, la nueva muestra será {0, 1, 1, 0, 1, 1, 1, 1}. A partir de esta muestra estimaremos la proporción P y el total de la clase A de los valores pares de X en la población, así como los errores de muestreo correspondientes. Tenemos:

$$\hat{P} = \frac{\sum_{i=1}^{8} A_i}{n} = \frac{6}{8} = 0{,}75 \quad (75\%) \qquad \hat{A} = N \cdot \hat{P} = 100 \frac{6}{8} = 75$$

$$e = \hat{\sigma}(\hat{P}) = \sqrt{(1 - f) \frac{\hat{P}\hat{Q}}{n - 1}} = \sqrt{\left(1 - \frac{8}{100}\right) \frac{0{,}75 \cdot 0{,}25}{8 - 1}} = 0{,}0246$$

$$e = \hat{\sigma}(\hat{A}) = N \cdot \hat{\sigma}(\hat{P}) = 100 \cdot 0{,}0246 = 2{,}46$$

El tamaño de muestra necesario para estimar la proporción de edades pares en la población con un error relativo de muestreo $e_{r\alpha} = 0{,}06$ y un coeficiente de confianza del 95% será:

$$n = \frac{\lambda_{r\alpha}^2 NQ}{(N-1)Pe_{r\alpha}^2 + \lambda_{r\alpha}^2 Q} = \frac{1{,}96^2 \cdot 100 \cdot (1-0{,}75)}{(100-1)\cdot 0{,}75 \cdot 0{,}06^2 + 1{,}96^2 \cdot (1-0{,}75)} = 78{,}22$$

Vamos a realizar a continuación *para muestreo con reposición* el cálculo del tamaño de muestra necesario para que el error relativo de muestreo sea 0,06 al estimar la proporción de edades pares de la población con un coeficiente de confianza del 95%. Utilizamos:

$$n = \frac{\lambda_\alpha^2 C_X^2}{e_{r\alpha}^2} = \frac{\lambda_\alpha^2 \dfrac{Q}{P}}{e_{r\alpha}^2} = \frac{1{,}96^2 \dfrac{1-0{,}75}{0{,}75}}{0{,}06^2} = 355$$

luego el tamaño de muestra necesario será $n = 355$, que supera al tamaño poblacional. Eso se debe a lo bajo que es el error especificado a cometer. En este caso habrá que aumentar el error a cometer. No obstante, se ha comprobado que el tamaño de muestra necesario para estimar el mismo parámetro cometiendo el mismo error siempre es mayor en el muestreo con reposición, lo que indica que este tipo de muestreo es menos preciso que el muestreo sin reposición. Esto concuerda también con el hecho de que los errores de muestreo siempre son menores en el caso de sin reposición.

3.7. Una muestra irrestricta aleatoria de 600 habitantes procedente de una población de $N = 15.000$ presenta los siguientes datos para la variable X = número de visitas anuales a doctores especialistas:

$$\sum_{i=1}^{600} X_i = 2946 \quad \text{y} \quad \sum_{i=1}^{600} X_i^2 = 18694$$

Hallar intervalos de confianza al 95% para el total y la media por habitante anuales de visitas a doctores especialistas en la población admitiendo normalidad para la distribución de los estimadores. Tomando la muestra anterior como muestra piloto, ¿qué tamaño de muestra será necesario para cometer un error absoluto de muestreo de 1.000 unidades al estimar el total de visitas a doctores especialistas en la población? ¿Y para cometer un error relativo de muestreo del 15%?

El total de visitas a doctores especialistas en la población, su error y el intervalo de confianza al 95% se estiman como sigue:

$$\hat{X} = N \cdot \bar{x} = 15000 \cdot \frac{2946}{600} = 73650 \quad \hat{S}^2 = \frac{1}{n-1}\left[\sum_{i=1}^{20} X_i^2 - \left(\sum_{i=1}^{20} X_i\right)^2 \Big/ n\right] = 7{,}06$$

$$\hat{\sigma}(\hat{X}) = \sqrt{N^2(1-f)\frac{\hat{S}^2}{n}} = \sqrt{15000^2\left(1-\frac{600}{15000}\right)\frac{7{,}06}{600}} = 1594{,}239$$

$$IC(X) = \hat{X} \pm \lambda_\alpha \hat{\sigma}(\hat{X}) = 73650 \pm 1{,}96 \cdot 1594{,}239 = (70526,\ 76775)$$

La media de visitas por habitante a doctores especialistas en la población, su error y el intervalo de confianza al 95% se estiman como sigue:

$$\bar{x} = \frac{2946}{600} = 4{,}91 \qquad \hat{\sigma}(\bar{x}) = \sqrt{(1-f)\frac{\hat{S}^2}{n}} = \sqrt{\left(1-\frac{600}{15000}\right)\frac{7{,}06}{600}} = 0{,}106282$$

$$IC(\bar{x}) = \bar{x} \pm \lambda_\alpha \hat{\sigma}(\bar{x}) = 4{,}91 \pm 1{,}96 \cdot 0{,}106282 = (4{,}70168,\ 5{,}11831)$$

El tamaño de muestra necesario para cometer un error absoluto de muestreo de 1.000 unidades al estimar el total poblacional de *X*, se puede calcular despejando *n* en la fórmula de la desviación típica del estimador del total, de la forma siguiente:

$$1000^2 = 15000^2\left(1-\frac{n}{15000}\right)\frac{7{,}06}{n} \Rightarrow n = \frac{15000^2 \cdot 7{,}06}{1000^2 + 15000 \cdot 7{,}06} = 1437$$

El tamaño de muestra necesario para cometer un error relativo de muestreo del 15% al estimar el total poblacional de X puede hallarse como sigue:

$$n = \frac{NC_{1,x}^2}{Ne_r^2 + C_{1,x}^2} = \frac{N\dfrac{S^2}{\bar{X}^2}}{Ne_r^2 + \dfrac{S^2}{\bar{X}^2}} = \frac{15000\dfrac{7{,}06}{4{,}91^2}}{15000 \cdot 0{,}15^2 + \dfrac{7{,}06}{4{,}91^2}} = 13$$

Hemos utilizado un valor de $S^2 = 7{,}06$ porque la muestra de tamaño 600 con los datos dados en el enunciado del problema se utiliza como muestra piloto.

3.8. Un sector industrial de Estados Unidos tiene un censo de 1000 fábricas. Hallar el tamaño de muestra necesario (número de fábricas) para que, con un grado de confianza del 95%, la estimación de la producción total del sector quede dentro del 10% de su valor verdadero. Se utiliza muestreo irrestricto aleatorio y se sabe por una muestra piloto que el coeficiente de variación poblacional es 0,6.

$$P(|\hat{X} - X| \le 0{,}10X) = 0{,}95 \Leftrightarrow P(-0{,}10X \le \hat{X} - X \le 0{,}10X) = 0{,}95 \Leftrightarrow$$

$$P\left(\frac{-0{,}10X}{\sigma(\hat{X})} \le \frac{\hat{X}-X}{\sigma(\hat{X})} \le \frac{0{,}10X}{\sigma(\hat{X})}\right) = 0{,}95 \Leftrightarrow P\left(\frac{-0{,}10X}{\sigma(\hat{X})} \le N(0,1) \le \frac{0{,}10X}{\sigma(\hat{X})}\right) = 0{,}95$$

$$\Rightarrow \frac{0{,}10X}{\sigma(\hat{X})} = \lambda_\alpha \Rightarrow 0{,}10 = \lambda_\alpha \frac{\sigma(\hat{X})}{X} = \lambda_\alpha \frac{\sigma(\hat{X})}{E(\hat{X})} = \lambda_\alpha Cv(\hat{X}) = e_{r\alpha} \text{ con } \lambda_\alpha = 1{,}96$$

Por lo tanto, el problema se traduce en calcular el tamaño de muestra necesario para cometer un error relativo de muestreo de 0,051 al estimar la producción total.

$$n = \frac{\lambda_\alpha^2 NC_{1,x}^2}{Ne_{r\alpha}^2 + \lambda_\alpha^2 C_{1,x}^2} = \frac{\lambda_\alpha^2 N\dfrac{S^2}{\bar{X}^2}}{Ne_{r\alpha}^2 + \lambda_\alpha^2\dfrac{S^2}{\bar{X}^2}} = \frac{\dfrac{\lambda_\alpha^2 N^2}{N-1}\left(\dfrac{\sigma}{\bar{X}}\right)^2}{Ne_{r\alpha}^2 + \dfrac{\lambda_\alpha^2 N}{N-1}\left(\dfrac{\sigma}{\bar{X}}\right)^2} = \frac{\dfrac{\lambda_\alpha^2 N}{N-1}(CV)^2}{e_{r\alpha}^2 + \dfrac{\lambda_\alpha^2}{N-1}(CV)^2} = \frac{\dfrac{1{,}96^2 \cdot 1000}{999} \cdot 0{,}6^2}{0{,}1^2 + \dfrac{1{,}96^2}{999}.0{,}6^2} = 122$$

3.9. Los partidos de izquierdas desean obtener información rápida sobre el número total de concejales que obtuvieron en las últimas elecciones en los 300 municipios más pequeños de una región española. Para ello se eligieron 50 municipios, y se obtuvieron los siguientes resultados:

Número de concejales por municipio X_i	Número de municipios n_i
0	2
1	7
2	5
3	7
4	8
5	10
6	5
7	3
8	2
9	1

Se pide:

1) Estimar el número total de concejales que obtuvieron los partidos de izquierdas en las últimas elecciones en la región en los municipios más pequeños.
2) Si se hubiera querido un error de muestreo inferior a 150 concejales, ¿cuántos municipios habría sido necesario seleccionar?

Tenemos $N = 300$ y $n = 50$. Para estimar el total de concejales que obtuvieron los partidos de izquierdas se procede como sigue:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{196}{50} = 3{,}92 \Rightarrow \hat{X} = N\bar{x} = 300 * 3{,}92 = 1176 \text{ concejales}$$

Como no se especifica lo contrario, se supone que el muestreo es sin reposición, en cuyo caso el error del estimador anterior al 99% de confianza es:

$$\sigma(\hat{X}) = \lambda_\alpha \sqrt{N^2\left(1 - \frac{n}{N}\right)\frac{\hat{S}^2}{n}} = 2{,}575 * \sqrt{300(300-50)}\,\frac{\sqrt{4{,}8098}}{\sqrt{50}} = 218{,}7189$$

$$\hat{S}^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} x_i^2 n_i - \frac{\left(\sum x_i\right)^2}{n}\right] = \frac{1}{49}\left[1004 - \frac{196^2}{50}\right] = 4{,}8098$$

Por tanto, la estimación del número de concejales obtenidos en los 300 municipios más pequeños de esa región durante las pasadas elecciones es de 1178 concejales. El error de muestreo con un 99% de confianza ha resultado ser 218,7, que en términos relativos (de coeficiente de variación) es:

$$\hat{C}_v(\hat{X}) = \frac{218{,}7189}{1176} \cdot 100 = 18{,}59\%$$

Para estimar el total de concejales con un error de muestreo inferior a 150, el número de municipios que habría sido necesario seleccionar se calculará como:

$$n = \frac{N^2 \lambda_\alpha^2 S^2}{e_T^2 + N\lambda_\alpha^2 S^2} = \frac{300^2 * 2{,}575^2 * 4{,}8098}{150^2 + 300 * 2{,}575^2 * 4{,}8098} = 89{,}51 \approx 90 \text{ municipios}$$

3.10. Un prestamista se dispone a contabilizar deudas atrasadas de 10000 clientes. Necesita aproximar la deuda sin cobrar y para ello elige una muestra aleatoria de 36 clientes, los cuales adeudan en media 7500 euros con un error (cuasidesviación típica) de 3000 euros. Realizar una estimación por intervalos al 95% de la deuda sin cobrar. ¿Qué tamaño de muestra deberá seleccionarse para estimar la deuda pendiente con un error de muestreo inferior a 2500000 euros.

Sea X la variable que mide la deuda sin cobrar. Dicha deuda total se estimará mediante:

$$\hat{X} = N\overline{x} = 1000 * 7500 = 7500000 \text{ euros}$$

El error de muestreo será:

$$\sigma(\hat{X}) = \sqrt{N^2 (1 - \frac{n}{N}) \frac{\hat{S}^2}{n}} = \sqrt{1000^2 \left(1 - \frac{36}{1000}\right) \frac{3000^2}{36}} = 2764{,}8$$

El intervalo de confianza para el total poblacional será:

$$[\hat{X} - \lambda_\alpha \sigma(\hat{X}), \hat{X} + \lambda_\alpha \sigma(\hat{X})] = [75000 - 1{,}96(2764{,}8);\ 75000 + 1{,}96(2764{,}8)] = [65217655{,}9; 84782344{,}1]$$

Para estimar la deuda pendiente con un error inferior a 2500000 euros, se debe elegir una muestra de tamaño superior al valor siguiente:

$$n = \frac{N^2 \lambda_\alpha^2 \hat{S}^2}{e_T^2 + N\lambda_\alpha^2 \hat{S}^2} = \frac{10000^2 * 1{,}96^2 * 3000^2}{(2500000)^2 + 10000 * 1{,}96^2 * 3000^2} = 524{,}19 \approx 525$$

3.11. En un recinto ferial se desea estimar la cantidad X gastada por visitante en sus instalaciones. Para ello, de entre los 500 visitantes de un día determinado, se seleccionó una muestra aleatoria simple de 100 y a la salida del recinto ferial se les preguntó la cantidad en euros que habían gastado. Se obtuvieron los siguientes datos:

$$\sum_{i=1}^{100} X_i = 250 \qquad \sum_{i=1}^{100} X_i^2 = 649{,}75$$

Hallar un intervalo de confianza al 95% para la cantidad media gastada por persona en el recinto ferial. ¿A cuántas personas se debería haber preguntado para que, con la misma confianza, el error de la estimación anterior no superarse los 75 euros? ¿Cuántas personas deberían haber sido preguntadas si se hubiera deseado estimar la proporción de personas insatisfechas con los servicios prestados en el recinto ferial con un error del 10% y una confianza del 95%?

El intervalo de confianza para la media poblacional será:

$$I = \left[\bar{x} - \lambda_\alpha \sqrt{(1 - \frac{n}{N}) \frac{\hat{S}^2}{n}}; \quad \bar{x} - \lambda_\alpha \sqrt{(1 - \frac{n}{N}) \frac{\hat{S}^2}{n}} \right]$$

$$\bar{x} = \frac{1}{n} \sum_{i!}^{n} x_i = \frac{250}{100} = 2{,}50 \text{ euros}$$

$$\hat{S}^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_i \right)^2 \right] = \frac{1}{99} \left[649{,}75 - \frac{1}{100} (250)^2 \right] = 0{,}25$$

El intervalo de confianza para el gasto medio en euros por persona en la feria será:

$$I = \left[2{,}5 - 1{,}96 \sqrt{(1 - \frac{100}{500}) \frac{0{,}25}{100}}; \quad 2{,}5 - 1{,}96 \sqrt{(1 - \frac{100}{500}) \frac{0{,}25}{100}} \right] = \left[2{,}4123; \quad 2{,}5876 \right]$$

Para un error en la estimación de la media de 75 euros, el número de personas que será necesario entrevistar será:

$$n = \frac{\lambda_\alpha^2 N \hat{S}^2}{e_\mu^2 N + \lambda_\alpha^2 \hat{S}^2} = \frac{1{,}96^2 * 500 * 0{,}25}{0{,}075^2 * 500 + 1{,}96^2 * 0{,}24} = 127{,}2761 \approx 128$$

El número de personas que deberían haber sido preguntadas si se hubiera deseado estimar la proporción de personas insatisfechas con los servicios prestados en el recinto ferial con un error del 10% y una confianza del 95% sería el siguiente:

$$n = \frac{\lambda_\alpha^2 NPQ}{e_\alpha^2 (N-1) + \lambda_\alpha^2 PQ} = \frac{1{,}96^2 * 500 * 0{,}5 * 0{,}5}{0{,}10^2 * 499 + 1{,}96^2 * 0{,}5 * 0{,}5} = 80{,}7005 \approx 81 \text{ personas}$$

Como no se tiene información sobre el valor de P, se toma $P = 0{,}5$.

3.12. Para tomar la decisión de mantener un determinado libro como texto oficial de una asignatura, se pretende tomar una muestra aleatoria simple entre los 1250 profesores de una universidad y enviarles un cuestionario a través del cual manifiesten si son favorables a la renovación del libro como texto oficial.

1) ¿Cuál deberá ser el número apropiado de profesores encuestados de entre los 1250 para obtener una estimación sobre la proporción de profesores favorables a la renovación del libro de texto con un error de muestreo inferior al 12% y una confianza del 90%?

2) Si de la encuesta realizada el año anterior se sabe que la proporción de profesores favorables al mantenimiento del libro de texto estará entre el 75% y el 85%, ¿cuál debería ser en este caso el número apropiado de profesores encuestados del apartado anterior?

3) Si finalmente se decidió enviar cuestionarios a 100 profesores, de los cuales tan sólo 35 no se manifestaron favorables a la renovación del libro de texto, estimar la proporción del número apropiado de profesores encuestados de entre los 1250 para obtener una estimación.

El número apropiado de profesores a encuestar de entre los 1250 para obtener una estimación sobre la proporción de profesores favorables a la renovación del libro de texto con un error de muestreo inferior al 12% y una confianza del 90% será el siguiente:

$$n = \frac{\lambda_\alpha^2 NPQ}{e_\alpha^2 (N-1) + \lambda_\alpha^2 PQ} = \frac{1{,}645^2 * 1250 * 0{,}5 * 0{,}5}{0{,}12^2 * 1249 + 1{,}645^2 * 0{,}5 * 0{,}5} = 45{,}2968 \approx 46 \text{ profesores}$$

Se ha utlizado $P = 1/2$ porque no se tiene información sobre P.

Para el caso de que se estime que la proporción oscilará entre el 75% y el 85%, en la fórmula para obtener el tamaño muestral se utilizará $P = 0{,}75$, pues es el que proporciona mayor variabilidad entre los posibles. Ahora tenemos:

$$n = \frac{1{,}645^2 * 1250 * 0{,}75 * 0{,}25}{0{,}12^2 * 1249 + 1{,}645^2 * 0{,}75 * 0{,}25} = 34{,}2954 \approx 35 \text{ profesores}$$

En el último apartado, como el estimador puntual de la proporción poblacional es la proporción muestral, tenemos:

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} A_i$$ (A_i =1 si el profesor i-ésimo mantiene el libro y $A_i = 0$ en caso contrario)

Como sólo 35 profesores de los 100 deciden la no renovación del libro de texto, tenemos:

$$\hat{P} = \frac{65}{100} = 0{,}65$$

El error de muestreo será:

$$e_p = \lambda_\alpha \sqrt{\frac{N-n}{N-1}\frac{pq}{n}} = 1{,}645\sqrt{\frac{1250-100}{1249} * \frac{0{,}65 * 0{,}35}{100}} = 0{,}0753$$

3.13. Una empresa industrial está interesada en el tiempo por semana que los científicos emplean para ciertas tareas triviales. Las hojas de control del tiempo de una muestra irrestricta aleatoria de $n = 50$ empleados muestran que la cantidad promedio de tiempo empleado en esas tareas es de 10,31 horas, con una varianza muestral de $S^2 = 2{,}25$. La compañía emplea N = 750 científicos. Estimar el número total de horas-hombre que se pierden por semana en las tareas insignificantes y establecer un límite para el error de estimación al 95% ($\lambda_\alpha = 2$).

Sea X el total de horas-hombre que se pierden por semana. Tenemos:

$$\hat{X} = N\bar{x} = 750(10{,}31) = 7732{,}5 \text{ horas}$$

Un límite para el error de estimación será el radio del intervalo de confianza al 95%:

$$\lambda_\alpha \sigma(\hat{X}) = 2\sqrt{N^2\left(1 - \frac{n}{N}\right)\frac{\hat{S}^2}{n}} = 2\sqrt{700^2\left(1 - \frac{50}{750}\right)\frac{2{,}25}{50}} = 307{,}4 \text{ horas}$$

3.14. Una muestra irrestricta aleatoria de $n = 100$ estudiantes del último año de un colegio fue seleccionada para estimar: (1) la fracción de entre los $N = 300$ estudiantes del último año que asistirán a una universidad, y (2) la fracción de estudiantes que han tenido trabajos de tiempo parcial durante su estancia en el colegio. Sean Y_i y X_i ($i = 1, 2, \ldots, 100$) las respuestas del i-ésimo estudiante seleccionado. Estableceremos que $Y_i = 0$ si el i-ésimo estudiante no planifica asistir a una institución superior, e $Y_i = 1$ si lo planifica. Asimismo, sea $X_i = 0$ si el estudiante i-ésimo no ha tenido trabajo durante su estancia en el colegio y sea $X_i = 1$ si lo ha tenido. Usando los datos de la muestra presentados en la tabla adjunta, estime P_1, la proporción de estudiantes del último año que planea asistir a una universidad y P_2, la proporción de estudiantes del último año que ha tenido un trabajo de tiempo parcial durante sus cursos en el colegio (incluyendo los veranos).

Estudiante	Y	X
1	1	0
2	0	1
3	0	1
4	1	1
5	0	0
6	0	0
7	0	1
.	.	.
.	.	.
96	0	1
97	1	0
98	0	1
99	0	1
100	1	1
	$\sum_{i=1}^{100} Y_i = 15$	$\sum_{i=1}^{100} X_i = 65$

Las estimaciones de las respectivas proporciones estarán dadas por las proporciones muestrales:

$$\hat{P}_1 = \frac{1}{100}\sum_{i=1}^{100} Y_i = \frac{15}{100} = 0{,}15 \qquad \hat{P}_2 = \frac{1}{100}\sum_{i=1}^{100} X_i = \frac{65}{100} = 0{,}65$$

Los límites para los respectivos errores de estimación al 95% estarán dados por los radios de los dos intervalos de confianza, que se calculan como sigue:

$$\lambda_\alpha \sigma(\hat{P}_1) = 2\sqrt{(1 - \frac{n}{N})\frac{\hat{P}_1\hat{Q}_1}{n-1}} = 2\sqrt{\left(1 - \frac{100}{300}\right)\frac{0{,}15 \cdot 0{,}85}{99}} = 0{,}059$$

$$\lambda_\alpha \sigma(\hat{P}_2) = 2\sqrt{(1 - \frac{n}{N})\frac{\hat{P}_2\hat{Q}_2}{n-1}} = 2\sqrt{\left(1 - \frac{100}{300}\right)\frac{0{,}65 \cdot 0{,}35}{99}} = 0{,}078$$

Hemos obtenido que el 15% de los estudiantes de último año planifica asistir a la universidad con un límite del error de la estimación del 5,9%, y el 65% de los estudiantes de último año ha tenido un trabajo a tiempo parcial durante su estancia en el colegio con un límite para el error de la estimación del 7,8%.

3.15. Mediante muestreo irrestricto aleatorio se obtiene una muestra de 50 trabajadores procedente de una población de 750 empleados de una multinacional. Al medir el salario mensual X en cientos de euros que perciben los trabajadores de la muestra se obtienen los siguientes datos:

$$\sum_{i=1}^{50} X_i = 454 \quad \text{y} \quad \sum_{i=1}^{50} X_i^2 = 4306$$

De esta muestra 20 trabajadores pertenecen al sector financiero de la multinacional, y al medir los salarios mensuales X sobre estos 20 empleados se obtienen los siguientes resultados:

$$\sum_{i=1}^{20} X_i = 172 \quad \text{y} \quad \sum_{i=1}^{20} X_i^2 = 1536$$

1º Estimar el salario medio mensual por trabajador y el total mensual de pagos en salarios de la multinacional para todos sus empleados y para los empleados del sector financiero, así como sus errores absolutos y relativos de muestreo.

2º Responder a las preguntas del apartado anterior para muestreo aleatorio simple con reposición comentando resultados y comparándolos con los del apartado 1.

Consideramos como población todos los empleados de la multinacional y como subpoblación todos los empleados del sector financiero de la multinacional.

Para estimar la media y el total de la población con $n = 50$ y $N = 750$ se tiene:

$$\bar{x} = \frac{\sum_{i=1}^{50} X_i}{n} = \frac{454}{50} = 9{,}08 \quad \text{y} \quad \hat{X} = N\bar{x} = 750\frac{\sum_{i=1}^{50} X_i}{n} = 750 \cdot 9{,}08 = 6810$$

Las estimaciones de los errores de muestreo serán:

$$\hat{V}(\bar{x}) = (1 - \frac{50}{750}) \frac{\frac{1}{49}\left[\underbrace{\sum_{i=1}^{50} X_i^2}_{4306} - \left(\underbrace{\sum_{i=1}^{50} X_i}_{454}\right)^2 \Big/ 50\right]}{50} = 0{,}07 \Rightarrow \hat{\sigma}(\bar{x}) = \sqrt{0{,}07} = 0{,}26$$

$$\hat{V}(\hat{X}) = N^2 \hat{V}(\bar{x}) = 750^2 \cdot 0{,}07 = 39375 \Rightarrow \hat{\sigma}(\hat{X}) = \sqrt{39375} = 198{,}43$$

Las estimaciones de los errores relativos de muestreo (coeficientes de variación de los estimadores) serán las siguientes:

$$\hat{Cv}(\bar{x}) = \frac{\hat{\sigma}(\bar{x})}{\bar{x}} = \frac{0{,}27}{9{,}08} = 0{,}029 \quad (2{,}9\%) \quad \text{y} \quad \hat{Cv}(\hat{X}) = \frac{\hat{\sigma}(\hat{X})}{\hat{X}} = \frac{198{,}43}{6810} = 0{,}029 \quad (2{,}9\%)$$

Evidentemente, los errores relativos de muestreo coinciden al estimar la media y el total para la población.

Hemos estimado que el salario medio de todos los trabajadores de la multinacional es de 908 euros mensuales y que los pagos totales mensuales de la multinacional en salarios de todos sus empleados es 681000 euros. Estas estimaciones tiene un error inferior al 3% (2,9%), lo que indica que son muy aceptables.

Para estimar la media y el total de la subpoblación con $n = 50$, $N = 750$, $n_1 = 20$ y N_1 desconocido, se tiene:

$$\bar{x}_1 = \frac{\sum_{i=1}^{20} X_i}{n_1} = \frac{172}{20} = 8{,}6 \quad \text{y} \quad \hat{X}_1 = N \cdot \frac{x_1}{n} = 750.\frac{\sum_{i=1}^{20} X_i}{50} = 750 \cdot \frac{172}{50} = 2580$$

$$\hat{V}(\bar{x}_1) = (1 - \frac{n}{N}) \frac{\frac{1}{n_1 - 1}\left[\sum_{i=1}^{20} X_i^2 - \left(\sum_{i=1}^{20} X_i\right)^2 \Big/ n_1\right]}{n_1} = (1 - \frac{50}{750}) \frac{\frac{1}{19}\left[1536 - 172^2/20\right]}{20} = 0{,}14$$

$$\hat{V}(\hat{X}_1) = N^2 (1 - \frac{n}{N}) \frac{\frac{1}{n - 1}\left[\sum_{i=1}^{20} X_i^2 - \left(\sum_{i=1}^{20} X_i\right)^2 \Big/ n\right]}{n} = 750^2 (1 - \frac{50}{750}) \frac{\frac{1}{49}\left[1536 - 172^2/50\right]}{50}$$

$$= 202354{,}28$$

Luego las estimaciones de los errores de muestreo para la subpoblación serán:

$$\hat{\sigma}(\bar{x}_1) = \sqrt{\hat{V}(\bar{x}_1)} = \sqrt{0{,}14} = 0{,}374 \quad \text{y} \quad \hat{\sigma}(\hat{X}_1) = \sqrt{\hat{V}(\hat{X}_1)} = \sqrt{202354{,}28} = 450$$

Las estimaciones de los errores relativos de muestreo (coeficientes de variación) para la subpoblación serán:

$$\hat{C}v(\bar{x}_1) = \frac{\hat{\sigma}(\bar{x}_1)}{\bar{x}_1} = \frac{0{,}374}{8{,}6} = 0{,}043 \quad (4{,}3\%) \quad \text{y} \quad \hat{C}v(\hat{X}_1) = \frac{\hat{\sigma}(\hat{X}_1)}{\hat{X}_1} = \frac{450}{2580} = 0{,}1744 \quad (17{,}44\%)$$

Para la subpoblación ya no coinciden los errores relativos de muestreo al estimar la media y el total.

Hemos estimado que el salario medio de los trabajadores del sector financiero de la multinacional es de 860 euros mensuales (algo inferior a los 908 euros mensuales de media cuando se consideran todos los trabajadores) y que los pagos totales mensuales de la multinacional en salarios de sus empleados del sector financiero es 258000 euros. Estas estimaciones tienen unos errores del 4,3% y del 17,44%, respectivamente. Es mucho más precisa la estimación del salario medio de los empleados del sector financiero que la estimación de los pagos totales a empleados de dicho sector.

En el caso de muestreo con reposición los estimadores son los mismos (para la población y para la subpoblación). Los errores de muestreo para la población y la subpoblación serán:

$$\hat{\sigma}_{CR}(\bar{x}) = \sqrt{\hat{V}_{CR}(\bar{x})} = \sqrt{\frac{\hat{V}(\bar{x})}{1-f}} = \frac{\sqrt{0,07}}{1-50/750} = 0,289$$

$$\hat{\sigma}_{CR}(\hat{X}) = \sqrt{\hat{V}_{CR}(\hat{X})} = \sqrt{\frac{\hat{V}(\hat{X})}{1-f}} = \frac{\sqrt{39375}}{1-50/750} = 212,28$$

$$\hat{\sigma}_{CR}(\bar{x}_1) = \sqrt{\hat{V}_{CR}(\bar{x}_1)} = \sqrt{\frac{\hat{V}(\bar{x}_1)}{1-f}} = \frac{\sqrt{0,14}}{1-50/750} = 0,4$$

$$\hat{\sigma}_{CR}(\hat{X}_1) = \sqrt{\hat{V}_{CR}(\hat{X}_1)} = \sqrt{\frac{\hat{V}(\hat{X}_1)}{1-f}} = \frac{\sqrt{202354,28}}{1-50/750} = 482,14$$

Se observa que los errores de muestreo al estimar la media y el total, tanto para la población como para la subpoblación, son mayores en el caso de muestreo con reposición que en el caso de muestreo sin reposición.

Las estimaciones de los errores relativos de muestreo (coeficientes de variación) para la población y la subpoblación serán:

$$\hat{Cv}(\bar{x}) = \frac{\hat{\sigma}_{CR}(\bar{x})}{\bar{x}} = \frac{0,289}{9,08} = 0,031 \quad (3,1\%) \quad \text{y} \quad \hat{Cv}(\hat{X}) = \frac{\hat{\sigma}_{CR}(\hat{X})}{\hat{X}} = \frac{212,28}{6810} = 0,031 \quad (3,1\%)$$

$$\hat{Cv}(\bar{x}_1) = \frac{\hat{\sigma}_{CR}(\bar{x}_1)}{\bar{x}_1} = \frac{0,4}{8,6} = 0,046 \quad (4,6\%) \quad \text{y} \quad \hat{Cv}(\hat{X}_1) = \frac{\hat{\sigma}_{CR}(\hat{X}_1)}{\hat{X}_1} = \frac{482,1}{2580} = 0,186 \quad (18,6\%)$$

Los errores relativos de muestreo al estimar la media y el total también son mayores en el caso de muestreo con reposición, tanto para la población como para la subpoblación.

3.16. La tabla adjunta muestra la distribución de frecuencias del número de residentes en cada una de las 197 ciudades de Estados Unidos que tenían más de 50000 habitantes en 1940.

Nº de residentes en miles de habitantes (clases)	Frecuencias absolutas	Nº de residentes en miles de habitantes (clases)	Frecuencias absolutas
50 - 100	105	650 -700	2
100 - 150	36	700 - 750	0
150 - 200	13	750 - 800	1
200 - 250	6	800 - 850	1
250 - 300	7	850 - 900	2
300 - 350	8	900 - 950	0
350 - 400	4	950 - 1000	0
400 - 450	1	1000 - 1050	0
450 - 500	3	1500 - 1550	1
500 - 550	0	1600 - 1650	1
550 - 600	2	1900 - 1950	1
600 - 650	1	3350 - 3400	1
		7450 - 7500	1

Calcular los errores absoluto y relativo de muestreo del número total de habitantes estimado en las 197 ciudades utilizando los siguientes métodos de muestreo:

1º) Muestro irrestricto aleatorio con tamaño de muestra $n = 50$.

2º) Muestreo que consiste en seleccionar las cinco ciudades más grandes y posteriormente una muestra irrestricta aleatoria de tamaño 45 para las 192 ciudades restantes.

Comenzaremos calculando la cuasivarianza para la distribución de frecuencias dada relativa a los tamaños de las ciudades. Considerando las marcas de clase se tiene:

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{197} n_i X_i^2 - \left(\sum_{i=1}^{197} n_i X_i\right)^2 \Big/ N\right] = \frac{1}{197-1}\left[85363125 - (46275)^2 / 197\right] = 380067{,}33$$

El error de muestreo para una muestra aleatoria simple sin reposición de tamaño 50 es:

$$\sigma(\hat{X}) = \sqrt{V(\hat{X})} = \sqrt{N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}} = \sqrt{197^2\left(1-\frac{50}{197}\right)\frac{380067{,}33}{50}} = 14836{,}79 \text{ miles de personas}$$

$$\text{Como } X = \sum_{i=1}^{197} n_i X_i = 46275 \Rightarrow Cv(\hat{X}) = \frac{14836{,}79}{46275} * 100 = 32\% \text{ (error relativo).}$$

En el segundo apartado consideramos la subpoblación de las cinco ciudades mayores (últimos cinco elementos de la tabla de frecuencias) que no presenta variabilidad porque se eligen todos sus elementos para la muestra, y en la subpoblación de las 192 ciudades restantes elegimos una muestra de tamaño 45. En esta última subpoblación calcularemos el error de muestreo ($N_1 = 192$ $n_1 = 45$).

$$S_1^2 = \frac{1}{N_1-1}\left[\sum_{i=1}^{192} n_i X_i^2 - \left(\sum_{i=1}^{192} n_i X_i\right)^2 \Big/ N_1\right] = \frac{1}{192-1}\left[9425000 - (30350)^2 / 192\right] = 24227{,}68$$

$$\sigma(\hat{X}) = \sqrt{V(\hat{X})} = \sqrt{N^2\left(1-\frac{n_1}{N_1}\right)\frac{S_1^2}{n_1}} = \sqrt{192^2\left(1-\frac{45}{192}\right)\frac{24227{,}268}{45}} = 3898{,}09 \text{ miles de personas}$$

$$\text{Como } X_1 = \sum_{i=1}^{192} n_i X_i = 30350 \Rightarrow Cv(\hat{X}) = \frac{3898{,}09}{30350} * 100 = 12{,}84\% \text{ (error relativo).}$$

3.17. Dos dentistas A y B hicieron una encuesta para investigar el estado de los dientes de 200 niños. El doctor A seleccionó una muestra irrestricta aleatoria de 20 niños y contó el número de dientes con caries de cada niño, con los siguientes resultados:

Nº de dientes con caries por niño	0	1	2	3	4	5	6	7	8	9	10
Nº de niños	8	4	2	2	1	1	0	0	0	1	1

El doctor B, utilizando las mismas técnicas dentales, examinó a los 200 niños y sólo registró aquellos que no tenían caries, encontrando que 60 niños no tenían dientes dañados.

1) Estudiar qué doctor obtiene estimaciones más precisas del número total de dientes con caries en los niños cuantificando la ganancia en precisión.

2) Realizar las estimaciones anteriores mediante intervalos de confianza al 95%. Comentar los resultados comparándolos con los del apartado anterior.

Para el doctor *A*, la estimación del número de dientes con caries será:

$$\hat{X} = N\bar{x} = 200\frac{0\cdot 8 + 1\cdot 4 + \cdots + 10\cdot 1}{20} = 200\cdot 2{,}1 = 420 \text{ dientes con caries.}$$

El error de muestreo de esta estimación es:

$$\hat{\sigma}(\hat{X}) = \sqrt{\hat{V}(\hat{X})} = \sqrt{N^2(1-\frac{n}{N})\frac{\hat{S}^2}{n}} = \sqrt{200^2(1-\frac{20}{200})\frac{8{,}62}{20}} = 123{,}04$$

$$\hat{S}^2 = \frac{1}{20-1}\left[\sum_{i=0}^{10} n_i X_i^2 - \left(\sum_{i=1}^{10} n_i X_i\right)^2 \Big/ n\right] = \frac{1}{19}\left[252 - (42)^2/20\right] = 8{,}62$$

La estimación por intervalos al 95% es $IC(X) = \hat{X} \pm \lambda_\alpha \hat{\sigma}(\hat{X}) = 420 \pm 1{,}96\cdot 123{,}04$.

Para el doctor B se considera la subpoblación de los 140 niños con caries resultante de eliminar de los 200 niños iniciales los 60 que no tenían caries. En cuanto a la muestra, hay que eliminar de la distribución inicial los ocho niños que tienen cero caries (20−8=12). La distribución muestral de frecuencias de esta subpoblación queda como sigue:

Nº de dientes con caries por niño	1	2	3	4	5	6	7	8	9	10
Nº de niños	4	2	2	1	1	0	0	0	1	1

Tenemos entonces $N_1 = 140$ y $n_1 = 12$.

$$\hat{X}_1 = N_1\bar{x}_1 = 140\frac{1\cdot 4 + \cdots + 10\cdot 1}{12} = 140\cdot 3{,}5 = 490 \text{ dientes con caries.}$$

El error de muestreo de esta estimación es:

$$\hat{\sigma}(\hat{X}_1) = \sqrt{\hat{V}(\hat{X}_1)} = \sqrt{N_1^2(1-\frac{n_1}{N_1})\frac{\hat{S}_1^{\,2}}{n_1}} = \sqrt{140^2(1-\frac{12}{140})\frac{9{,}545}{12}} = 419{,}370$$

$$\hat{S}_1^2 = \frac{1}{12-1}\left[\sum_{i=1}^{10} n_i X_i^2 - \left(\sum_{i=1}^{10} n_i X_i\right)^2 \Big/ n_1\right] = \frac{1}{11}\left[252 - (42)^2/12\right] = 9{,}545$$

Se observa que la precisión del doctor B es bastante menor (error mayor).

La estimación por intervalos al 95% es $IC(X_1) = \hat{X}_1 \pm \lambda_\alpha \hat{\sigma}(\hat{X}_1) = 490 \pm 1{,}96\cdot 419{,}37$.

EJERCICIOS PROPUESTOS

3.1. Consideramos una población finita de seis elementos sobre los que medimos una variable X, obteniendo como resultados $X_i = \{8, 3, 1, 11, 4, 7\}$, i = 1, ..., 6. Mediante muestreo irrestricto aleatorio se extraen muestras de tamaño 2. Se pide:

1) ¿Cuántos elementos tiene el espacio muestral? Especificar dicho espacio muestral y las probabilidades asociadas a las muestras. Hallar las distribuciones en el muestreo de los estimadores de la media y del total de X, así como de los estimadores de sus varianzas. Comprobar la insesgadez de los estimadores y que se cumple $V(\bar{x}) = (1-f)\frac{S^2}{n}$, $V(\hat{X}) = N^2(1-f)\frac{S^2}{n}$ y $E(\hat{S}^2) = S^2$, así como que el estimador T = Total muestral no es insesgado del total poblacional X.

2) Hallar el tamaño de muestra necesario para que el error de muestreo sea 2 al estimar la media de la población. ¿Y al estimar el total poblacional? Hallar también el tamaño de muestra necesario para que el error relativo de muestreo sea 0.48 en las mismas estimaciones. Calcular todos los tamaños de muestra anteriores en presencia de un coeficiente de confianza adicional del 95%. Comentar los resultados.

3) Contestar a todas las preguntas del apartado anterior para muestreo con reposición. Comparar los resultados con los de muestreo sin reposición. Comentar los resultados.

4) ¿A partir de qué tamaño poblacional N el aumento del tamaño muestral n no interviene en el error absoluto de muestreo para la estimación de la media? ¿Cuánto valdrá N con un coeficiente de confianza del 95%? Hallar intervalos de confianza al 95% para la media y el total basados en las muestras de elementos pares. Si al medir una variable X sobre los elementos de la población se obtienen los valores {1, 3, 4}, ¿cuál de todos los métodos de muestreo es más preciso al estimar el total poblacional mediante un estimador lineal insesgado apropiado?

3.2. Mediante muestreo irrestricto aleatorio se trata de estimar la proporción y el total de piezas correctas producidas en un proceso industrial en el que se fabrican un total de 6000 unidades. Una muestra piloto ha suministrado 1/3 de piezas defectuosas. Se pide:

1) Hallar el tamaño de muestra necesario para que el error de muestreo sea de una décima al estimar la proporción de piezas correctas producidas en el proceso industrial. Hallar también el tamaño de muestra necesario para que el error relativo de muestreo sea de 20% en la misma estimación.

2) Hallar el tamaño de muestra necesario para que el error de muestreo sea de 600 unidades al estimar el total de piezas correctas con un coeficiente de confianza del 99,7% y suponiendo muestreo aleatorio simple con reposición. Hallar dicho tamaño en las condiciones anteriores pero para un error relativo de muestreo del 10%.

3.3. Con el objetivo del análisis de la divisibilidad de un conjunto de números consideramos la población virtual $X_i = \{2, 13, 17, 23, 6, 1\}$, $i = 1, ..., 6$. Mediante muestreo irrestricto aleatorio se extraen muestras de tamaño 2.

1) Se trata de estimar los parámetros poblacionales PROPORCIÓN DE NÚMEROS PRIMOS y TOTAL DE NÚMEROS PRIMOS mediante estimadores insesgados basados en las muestras del espacio muestral. Hallar la distribución en el muestreo de dichos estimadores y de las estimaciones insesgadas de sus varianzas. Comprobar todas las insesgadeces y que se cumplen las relaciones $V(\hat{P}) = (1-f)\dfrac{\frac{N}{N-1}PQ}{n}$, $V(\hat{A}) = N^2(1-f)\dfrac{\frac{N}{N-1}PQ}{n}$ y $E(\hat{S}^2) = S^2$, así como que el estimador T = Total de números primos en las muestras no es insesgado del total de clase poblacional A.

3) Hallar el tamaño de muestra necesario para que el error de muestreo sea 1/4 al estimar la proporción de números primos de la población. Hallar también el tamaño de muestra necesario para que el error relativo de muestreo sea del 2% en la misma estimación.

4) Hallar intervalos de confianza al 99% ($\alpha = 0{,}01$) para el total y la proporción de números primos en la población basados en las muestras cuyos dos elementos son números no primos. Tenemos como dato conocido que $F^{-1}(0.995) = 2{,}57$, siendo F la función de distribución de la normal (0,1). Comentar los resultados.

5) Hallar el tamaño de muestra necesario para que el error de muestreo sea 6 al estimar el total de números primos de la población con un coeficiente de confianza del 99% y suponiendo muestreo aleatorio simple con reposición. Hallar dicho tamaño en las condiciones anteriores pero para un error relativo de muestreo del 90%. Comentar los resultados.

3.4. Un investigador está interesado en estimar la ganancia en peso total en 0 a 4 semanas de $N = 1000$ polluelos alimentados con una nueva ración. Obviamente, pesar cada ave sería tedioso y lento. Por lo tanto, determinar el número de polluelos que serán seleccionados en este estudio para estimar τ con un límite para el error de estimación igual a 1000 gramos. Muchos estudios similares sobre nutrición de polluelos se han llevado a cabo en el pasado. Usando los datos de esos estudios, el investigador encontró que σ^2, la varianza poblacional, fue aproximadamente igual a 36,00 gramos. Determine el tamaño de muestra requerido.

3.5. Una muestra irrestricta aleatoria de $n = 100$ medidores de agua es controlada dentro de una comunidad para estimar el promedio de consumo de agua diario por casa durante un periodo estacional seco. La media y la varianza muestrales fueron $\bar{y} = 12{,}5$ y $s^2 = 1252$. Si suponemos que hay $N = 10000$ casas dentro de la comunidad, estimar μ, el promedio de consumo diario verdadero, y establezca un límite para el error de estimación.

CAPÍTULO

4

MUESTREO ESTRATIFICADO SIN Y CON REPOSICIÓN

OBJETIVOS

1. Presentar el concepto de muestreo estratificado.
2. Comprender las especificaciones del muestreo estratificado.
3. Analizar los estimadores y sus errores en muestreo aleatorio estratificado sin reposición.
4. Estimar los errores en muestreo aleatorio estratificado sin reposición.
5. Analizar los estimadores y sus errores en muestreo estratificado con reposición.
6. Estimar los errores en muestreo aleatorio estratificado sin reposición.
7. Comprender el concepto de afijación de la muestra.
8. Estudiar los distintos tipos de afijación.
9. Especificar los errores de los estimadores en función de los distintos tipos de afijación.
10. Analizar el tamaño de la muestra en general.
11. Estudiar el tamaño de la muestra en función de los distintos tipos de afijación.
12. Comparar la eficiencia de los distintos tipos de afijación.
13. Presentar el concepto de postestratificación.
14. Analizar estimadores y errores en postestratifiacción.

ÍNDICE

1. Concepto de muestreo estratificado.

2. Muestreo estratificado sin reposición. Estimadores y errores.

3. Muestreo estratificado con reposición. Estimadores y errores.

4. Afijación de la muestra. Tipos de afijación y errores de los estimadores para muestreo sin reposición.

5. Afijación de la muestra. Tipos de afijación y errores de los estimadores para muestreo con reposición.

6. Tamaño de la muestra para muestreo sin reposición.

7. Tamaño de la muestra para muestreo con reposición.

8. Comparación de eficiencias en muestreo estratificado.

9. Postestratificación.

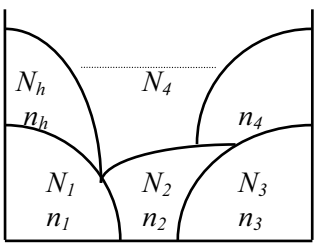
10. Problemas resueltos

11. Ejercicios propuestos

CONCEPTO DE MUESTREO ESTRATIFICADO

Supongamos que la población objeto de estudio, formada por N unidades elementales, se divide en L subpoblaciones o estratos, los cuales constituyen una partición, es decir, no se solapan y la unión de todos ellos es el total. De forma más precisa podemos decir que en el muestreo estratificado, una *población heterogénea* con N unidades $\{u_i\}_{\,i=1,2,\ldots,N}$ se subdivide en L *subpoblaciones disjuntas lo más homogéneas posible* (*que forman una partición*) de tamaños $N_1, N_2, \ldots, N_L$., denominadas *estratos* $\{u_{hi}\}_{\substack{h=1,2,\cdots,L \\ i=1,2,\cdots N_h}}$.

El muestreo estratificado es un tipo de muestreo de unidades elementales ya que la muestra estratificada de tamaño n se obtiene seleccionando n_h elementos ($h = 1, 2, \ldots, L$) de cada uno de los L estratos en los que se subdivide la población de forma independiente. Si la selección en cada estrato es aleatoria simple y de forma independiente, el muestreo se *denomina muestreo aleatorio estratificado*, pero en general nada impide utilizar diferentes tipos de selección en cada estrato. Si el muestreo aleatorio en cada estrato es sin reposición, el muestreo estratificado es sin reposición, y si el muestreo aleatorio en cada estrato es con reposición, el muestreo estratificado es con reposición. El gráfico siguiente muestra la población dividida en h estratos de tamaño N_h, en cada de los cuales elegimos de modo independiente n_h unidades (por muestreo aleatorio simple si no se especifica otra cosa) para la muestra estratificada de tamaño n..



POBLACIÓN

A continuación se expresa de modo esquemático la formación de estratos en la población y la formación de la muestra estratificada de la forma siguiente:

POBLACIÓN

$$\{u_1 u_2 \cdots u_N\} \xrightarrow{\textit{Se divide en L estratos}} \left\{ \begin{array}{l} u_{11} u_{12} \cdots u_{1N_1} \\ u_{21} u_{22} \cdots u_{2N_2} \\ \cdots\cdots\cdots\cdots\cdots \\ u_{L1} u_{L2} \cdots u_{LN_L} \end{array} \right\} \sum_{h=1}^{L} N_h = N$$

MUESTRA

$$\{u_1 u_2 \cdots u_n\} \xrightarrow{\textit{Se extrae en cada estrato}} \left\{ \begin{array}{l} u_{11} u_{12} \cdots u_{1n_1} \\ u_{21} u_{22} \cdots u_{2n_2} \\ \cdots\cdots\cdots\cdots\cdots \\ u_{L1} u_{L2} \cdots u_{Ln_L} \end{array} \right\} \sum_{h=1}^{L} n_h = n$$

El principal objetivo del muestreo estratificado es mejorar la precisión de las estimaciones reduciendo los errores de muestreo. Intenta minimizar la varianza de los estimadores mediante la creación de estratos lo más homogéneos posible entre sus elementos (para que los estimadores del estrato sean precisos) y lo más heterogéneos entre sí (para tener el máximo de información). Otros objetivos del muestreo estratificado son los siguientes:

1. Obtener estimaciones separadas para cada uno de los estratos.

2. Hacer un uso más racional de la organización administrativa.

3. Paliar los defectos del marco, aislando esos defectos en algunos estratos.

Es muy conveniente utilizar muestreo estratificado cuando existe una variable precisa para la estratificación cuyos valores permitan dividir convenientemente la población en estratos homogéneos. Las variables utilizadas para la estratificación deberán estar correlacionadas con las variables objeto de la investigación. Por ejemplo, para realizar estadísticas sobre los ingresos de las familias en una ciudad puede estratificarse según los valores de la variable cualificación profesional de los cabezas de sus componentes (a más cualificación normalmente hay más ingresos, con lo que los estratos resultarán homogéneos). Si se quiere estudiar el volumen de negocio de los establecimientos de venta al público de una ciudad, se puede utilizar como variable de estratificación su número de empleados, y clasificar (estratificar) los establecimientos en grandes superficies, supermercados, tiendas grandes, tiendas pequeñas y otros, según el número de empleados; así resulta una división de los establecimientos en grupos homogéneos. Si se quiere estudiar características de hospitales se puede utilizar la variable de estratificación número de pacientes, para estratificarlos en grandes hospitales, clínicas medias y clínicas pequeñas, resultando así grupos de hospitales con problemática similar. Para realizar estadísticas en el sector educativo puede utilizarse la variable de estratificación nivel de enseñanza, tomando como estratos los niveles de enseñanza infantil, enseñanza primaria, enseñanza secundaria obligatoria, bachillerato y enseñanza universitaria (cada estrato tiene así unas características muy peculiares que lo hacen homogéneo).

MUESTREO ESTRATIFICADO SIN REPOSICIÓN: ESTIMADORES Y ERRORES

En muestreo estratificado un parámetro poblacional puede escribirse como $\theta = \sum_{h}^{L} \sum_{i}^{N_h} Y_{hi}$.

El parámetro θ puede ser estimado mediante la suma extendida a todos los estratos de los estimadores lineales insesgados de Horvitz y Thompson en cada estrato, es decir, mediante:

$$\hat{\theta} = \sum_{h}^{L} \sum_{i}^{n_h} \frac{Y_{hi}}{\pi_{hi}}$$

donde π_{hi} es la probabilidad de que la unidad u_{hi} pertenezca a la muestra $(\tilde{X}_h)$ de n_h unidades, obtenida de entre las N_h unidades del estrato h-ésimo. Para los diferentes estimadores tendremos las siguientes expresiones:

$$\theta = X \Rightarrow Y_{hi} = X_{hi} \Rightarrow \hat{X}_{st} = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{X_{hi}}{\pi_{hi}} = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{X_{hi}}{n_h/N_h} = \sum_{h=1}^{L}N_h\underbrace{\frac{1}{n_h}\sum_{i=1}^{n_h}X_{hi}}_{\hat{\bar{X}}_h=\bar{x}_h} = \sum_{h=1}^{L}N_h\bar{x}_h = \sum_{h=1}^{L}\hat{X}_h$$

$$\theta = \bar{X} \Rightarrow Y_{hi} = \frac{X_{hi}}{N} \Rightarrow \hat{\bar{X}}_{st} = \bar{x}_{st} = \sum_{h=1}^{L}\frac{1}{N}\sum_{i=1}^{n_h}\frac{X_{hi}}{\pi_{hi}} = \sum_{h=1}^{L}\frac{1}{N}\sum_{i=1}^{n_h}\frac{X_{hi}}{n_h/N_h} = \sum_{h=1}^{L}\underbrace{\frac{N_h}{N}}_{W_h}\frac{1}{n_h}\sum_{i=1}^{n_h}X_{hi} = \sum_{h=1}^{L}W_h\bar{x}_h$$

$$\theta = A \Rightarrow Y_{hi} = A_{hi} \Rightarrow \hat{A}_{st} = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{A_{hi}}{\pi_{hi}} = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{A_{hi}}{n_h/N_h} = \sum_{h=1}^{L}N_h\underbrace{\frac{1}{n_h}\sum_{i=1}^{n_h}A_{hi}}_{\hat{P}_h} = \sum_{h=1}^{L}N_h\hat{P}_h = \sum_{h=1}^{L}\hat{A}_h$$

$$\theta = P \Rightarrow Y_{hi} = \frac{A_{hi}}{N} \Rightarrow \hat{P}_{st} = \sum_{h=1}^{L}\frac{1}{N}\sum_{i=1}^{n_h}\frac{A_{hi}}{\pi_{hi}} = \sum_{h=1}^{L}\frac{1}{N}\sum_{i=1}^{n_h}\frac{A_{hi}}{n_h/N_h} = \sum_{h=1}^{L}\underbrace{\frac{N_h}{N}}_{W_h}\underbrace{\frac{1}{n_h}\sum_{i=1}^{n_h}A_{hi}}_{\hat{P}_h} = \sum_{h=1}^{L}W_h\hat{P}_h$$

El estimador del total poblacional en muestreo estratificado aleatorio es la suma de los estimadores del total en cada estrato y los factores de elevación son N_h / n_h. El estimador de la media en muestreo estratificado aleatorio es la media ponderada de los estimadores de la media en cada estrato, siendo los coeficientes de ponderación $W_h = N_h/N$ de suma unitaria, que a su vez son los factores de elevación. El estimador del total de clase en muestreo estratificado aleatorio es la suma de los estimadores del total de clase en cada estrato. El estimador de la proporción en muestreo estratificado aleatorio es la media ponderada de los estimadores de la proporción en cada estrato, siendo los coeficientes de ponderación $W_h = N_h/N$ de suma unitaria. Las varianzas de los estimadores y sus errores son ($f_h = n_h / N_h$):

$$V\left(\hat{X}_{st}\right) = \sum_{h=1}^{L}N_h^2\left(1-f_h\right)\frac{S_h^2}{n_h}, \quad V\left(\bar{x}_{st}\right) = V\left(\sum_{h=1}^{L}W_h\,\bar{x}_h\right) = \sum_{h=1}^{L}W_h^2\left(1-f_h\right)\frac{S_h^2}{n_h}$$

$$V\left(\hat{A}_{st}\right) = \sum_{h=1}^{L}N_h^2\left(1-f_h\right)\frac{N_h}{N_h-1}\frac{P_hQ_h}{n_h}, \quad V\left(\hat{P}_{st}\right) = \sum_{h=1}^{L}W_h^2\left(1-f_h\right)\frac{N_h}{N_h-1}\frac{P_hQ_h}{n_h}$$

$$\hat{V}\left(\hat{X}_{st}\right) = \sum_{h=1}^{L}N_h^2\left(1-f_h\right)\frac{\hat{S}_h^2}{n_h}, \quad \hat{V}\left(\bar{X}_{st}\right) = \sum_{h=1}^{L}W_h^2\left(1-f_h\right)\frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}\left(\hat{A}_{st}\right) = \sum_{h=1}^{L}N_h^2\left(1-f_h\right)\frac{\hat{P}_h\hat{Q}_h}{n_h-1}, \quad \hat{V}\left(\hat{P}_{st}\right) = \sum_{h=1}^{L}W_h^2\left(1-f_h\right)\frac{\hat{P}_h\hat{Q}_h}{n_h-1}$$

S_h^2 = cuasivarianza poblacional en el estrato h, $\hat{S}_h^2$ = cuasivarianza muestral en el estrato h.

MUESTREO ESTRATIFICADO CON REPOSICIÓN: ESTIMADORES Y ERRORES

Para el caso del muestreo estratificado con reposición los estimadores son los mismos, y sus varianzas son las siguientes:

$$V\left(\hat{X}_{st}\right) = \sum_{h=1}^{L}N_h^2\frac{\sigma_h^2}{n_h}, \quad V\left(\bar{x}_{st}\right) = \sum_{h=1}^{L}W_h^2\frac{\sigma_h^2}{n_h}, \quad V\left(\hat{A}_{st}\right) = \sum_{h=1}^{L}N_h^2\frac{P_hQ_h}{n_h}, \quad V\left(\hat{P}_{st}\right) = \sum_{h=1}^{L}W_h^2\frac{P_hQ_h}{n_h}$$

Las estimaciones de los errores (estimaciones de varianzas) son las siguientes:

$$\hat{V}\left(\hat{X}_{st}\right)=\sum_{h=1}^{L}N_h^2\frac{\hat{S}_h^2}{n_h},\quad \hat{V}\left(\overline{X}_{st}\right)=\sum_{h=1}^{L}W_h^2\frac{\hat{S}_h^2}{n_h},\quad \hat{V}\left(\hat{A}_{st}\right)=\sum_{h=1}^{L}N_h^2\frac{\hat{P}_h\hat{Q}_h}{n_h-1},\quad \hat{V}\left(\hat{P}_{st}\right)=\sum_{h=1}^{L}W_h^2\frac{\hat{P}_h\hat{Q}_h}{n_h-1}$$

AFIJACIÓN DE LA MUESTRA: TIPOS DE AFIJACIÓN Y ERRORES DE LOS ESTIMADORES PARA MUESTREO SIN REPOSICIÓN

Se llama afijación de la muestra al reparto, asignación, adjudicación, adscripción o distribución del tamaño muestral *n* entre los diferentes estratos; esto es, a la determinación de los valores de n_h que verifiquen $n_1 + n_2 + ... + n_L = n$. Pueden establecerse muchas afijaciones o maneras de repartir la muestra entre los estratos, pero las más importantes son: la afijación uniforme, la afijación proporcional, la afijación de varianza mínima y la afijación óptima.

Afijación uniforme

Consiste en asignar el mismo número de unidades muestrales a cada estrato, con lo que se tomarán todos los n_h iguales a n/L, aumentando o disminuyendo este tamaño en una unidad si *n* no fuese múltiplo de *L*, esto es, $n_h = E(n/L) + 1$, donde *E* denota la parte entera.

$$n_h = k\forall h = 1\cdots L\Rightarrow \sum_{h=1}^{L}n_h=\sum_{h=1}^{L}k\Rightarrow n=Lk\Rightarrow f_h=\frac{n_h}{N_h}=\frac{k}{N_h}$$

Para este tipo de afijación, las varianzas de los estimadores y sus estimaciones se hallan sustituyendo en las fórmulas generales f_h por k/N_h. Este tipo de afijación da la misma importancia a todos los estratos, en cuanto a tamaño de la muestra, con lo cual favorecerá a los estratos de menor tamaño y perjudicará a los grandes en cuanto a precisión. Sólo es conveniente en poblaciones con estratos de tamaño similar.

Afijación proporcional

Consiste en asignar a cada estrato un número de unidades muestrales proporcional a su tamaño. Las *n* unidades de la muestra se distribuyen proporcionalmente a los tamaños de los estratos expresados en número de unidades. Tenemos:

$$n_h = N_h k\Rightarrow \underbrace{\sum_{h=1}^{L}n_h}_{n}=\sum_{h=1}^{L}N_h k=k\underbrace{\sum_{h=1}^{L}N_h}_{N}\Rightarrow n=kN\Rightarrow k=\frac{n}{N}=f$$

$$f_h=\underbrace{\frac{n_h}{N_h}}_{\pi_{hi}}=\frac{N_h k}{N_h}=k=f\quad W_h=\frac{N_h}{N}=\frac{n_h/k}{n/k}=\frac{n_h}{n}$$

Para este tipo de afijación, las varianzas de los estimadores serán:

$$V\left(\hat{X}_{st}\right)=\frac{(1-k)}{k}\sum_{h=1}^{L}N_h\cdot S_h^2,\quad V\left(\bar{x}_{st}\right)=\frac{(1-k)}{n}\sum_{h=1}^{L}W_h\cdot S_h^2$$

$$V\left(\hat{A}_{st}\right)=\frac{(1-k)}{k}\sum_{h=1}^{L}\frac{N_h^2}{N_h-1}\cdot P_hQ_h,\quad V\left(\hat{P}_{st}\right)=\frac{(1-k)}{k}\sum_{h=1}^{L}\frac{N_h^2/N}{N_h-1}\cdot P_hQ_h$$

En afijación proporcional los estimadores de media y total pueden expresarse como sigue:

$$\hat{X}_{st}=\sum_{h=1}^{L}N_h\bar{x}_h=\sum_{h=1}^{L}\frac{n_h}{k}\bar{x}_h=\frac{1}{K}\sum_{h=1}^{L}n_h\underbrace{\bar{x}_h}_{x_h/n_h}=\frac{\sum_{h=1}^{L}x_h}{k}=\frac{x}{f}=\frac{Total\ muestral}{Fraccion\ demuestreo}$$

$$\hat{\bar{X}}_{st}=\bar{x}_{st}=\sum_{h=1}^{L}W_h\bar{x}_h=\sum_{h=1}^{L}\frac{n_h}{n}\bar{x}_h=\frac{1}{n}\sum_{h=1}^{L}n_h\underbrace{\bar{x}_h}_{x_h/n_h}=\frac{\sum_{h=1}^{L}x_h}{n}=\frac{Total\ muestral}{Tamaño\ de\ muestra}$$

A la vista de los resultados anteriores, en afijación proporcional, podemos asegurar lo siguiente:

- Las fracciones de muestreo en los estratos son iguales y coinciden con la fracción global de muestreo, siendo su valor la constante de proporcionalidad.

- Los coeficientes de ponderación W_h se obtienen exclusivamente a partir de la muestra, pues para su cálculo sólo son necesarios valores muestrales (n_h y n).

- El estimador insesgado para el total poblacional puede expresarse como el cociente entre el total muestral y la fracción de muestreo, o lo que es lo mismo, como el producto del total muestral por la inversa de la fracción de muestreo. Similar propiedad tiene el estimador insesgado para el total de clase (producto del total de clase muestral por la inversa de la fracción de muestreo).

- El estimador insesgado para la media poblacional puede expresarse como el cociente entre el total muestral y el tamaño de la muestra. Similar propiedad tiene el estimador insesgado para la proporción poblacional (cociente entre el total de clase muestral y el tamaño de la muestra).

- Como $\pi_{hi}=\frac{n_h}{N_h}=k=f,$ todas las unidades de la población tienen la misma probabilidad de figurar en la muestra de n unidades; es decir, estamos en el caso de muestras autoponderadas.

Afijación de mínima varianza (o afijación de Neyman)

La afijación de mínima varianza o afijación de Neyman consiste en determinar los valores de n_n (número de unidades que se extraen del estrato h-ésimo para la muestra) de forma que para un tamaño de muestra fijo igual a n la varianza de los estimadores sea mínima.

La expresión para n_h es $n_h = n \cdot \dfrac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} = n \cdot \dfrac{\frac{N_h}{N} S_h}{\sum_{h=1}^{L} \frac{N_h}{N} S_h} = n \cdot \dfrac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}$.

Vemos que los valores de n_h son proporcionales a los productos $N_h \cdot S_h$ y en el supuesto de que $S_h = S$, $\forall h = 1, 2, ..., L$ esta afijación de mínima varianza coincidiría con la proporcional, tal y como se ve a continuación:

$$S_h = S \Rightarrow n_h = n \cdot \frac{N_h S}{\sum_{h=1}^{L} N_h S} = \frac{n N_h}{N} = k N_h \ \ con\ k = \frac{n}{N}$$

La utilidad de esta afijación es mayor si hay grandes diferencias en la variabilidad de los estratos. En otro caso, la mayor sencillez y autoponderación de la afijación proporcional hacen preferible el empleo de ésta.

Una vez calculados los n_h para afijación de mínima varianza, vamos a ver cuánto vale la *varianza del estimador de la media y del total* para este tipo de afijación. Tenemos:

$$V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2, \quad V(\hat{X}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} N_h S_h^2$$

Si se quiere la afijación y la expresión de la *varianza mínima para el estimador de la proporción y el total de clase,* basta sustituir en la fórmula anterior S_h^2 por $P_h Q_h N_h/(N_h-1)$.

Afijación óptima

La afijación óptima consiste en determinar los valores de n_h (número de unidades que se extraen del estrato h-ésimo para la muestra) de forma que para un coste fijo C la varianza de los estimadores sea mínima. El coste fijo C será la suma de los costes derivados de la selección de las unidades muestrales de los estratos; es decir, si c_h es el coste por unidad de muestreo en el estrato h, el coste total de selección de las n_h unidades muestrales en ese estrato será $c_h n_h$. Sumando los costes $c_h n_h$ para los L estratos tenemos el coste total de selección de la muestra estratificada.

Podemos escribir que $n_h = n \cdot \dfrac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}} = n \cdot \dfrac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}}$

Vemos que los valores de n_h son proporcionales a los productos $N_h \cdot S_h / \sqrt{c_h}$ y en el supuesto de que $C_h = k$ $\forall h = 1, 2, ..., L$ (coste constante en todos los estratos) la afijación óptima coincide con la de mínima varianza, y si además $S_h = S$, $\forall h = 1, 2, ..., L$ la afijación óptima coincidirá con la de mínima varianza y con la proporcional.

Valor de la varianza mínima

Una vez calculados los n_h para afijación óptima, vamos a ver cuánto vale la *varianza del estimador de la media y del total* para este tipo de afijación. Tenemos:

$$V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right) - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2$$

$$V(\hat{X}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} N_h S_h \sqrt{c_h}\right) - \frac{1}{N}\sum_{h=1}^{L} N_h S_h^2$$

Si se quiere la afijación óptima y la expresión de la *varianza mínima para el estimador de la proporción y el total de clase,* basta sustituir en la fórmula anterior S_h^2 por $P_h Q_h N_h/(N_h\text{-}1)$.

AFIJACIÓN DE LA MUESTRA: TIPOS DE AFIJACIÓN Y ERRORES DE LOS ESTIMADORES PARA MUESTREO CON REPOSICIÓN

Dada la forma en que están definidos los cálculos de los n_h para las afijaciones uniforme y proporcional, dichas afijaciones no van a verse afectadas por el hecho de que el muestreo sea con o sin reposición. Sin embargo, sí variarán las varianzas de los estimadores. Las afijaciones de mínima varianza y óptima sí van a verse afectadas por la existencia de reposición o no, ya que el cálculo de n_h depende de las varianzas en los estratos.

Afijación uniforme

Para este tipo de afijación, las varianzas de los estimadores serán:

$$V(\hat{X}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{\sigma_h^2}{k}, \quad V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{k}, \quad V(\hat{A}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{P_h Q_h}{k}, \quad V(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{P_h Q_h}{k}$$

Afijación proporcional

Para este tipo de afijación las varianzas de los estimadores serán:

$$V(\hat{X}_{st}) = \frac{1}{k}\sum_{h=1}^{L} N_h \sigma_h^2, \quad V(\hat{A}_{st}) = \frac{1}{k}\sum_{h=1}^{L} N_h P_h Q_h, \quad V(\bar{x}_{st}) = \frac{1}{n}\sum_{h=1}^{L} W_h \sigma_h^2, \quad V(\hat{P}_{st}) = \frac{1}{n}\sum_{h=1}^{L} W_h \frac{P_h Q_h}{k}$$

Afijación de mínima varianza (o afijación de Neyman)

Tenemos:

$$n_h = n \cdot \frac{W_h \sigma_h}{\sum_{h=1}^{L} W_h \sigma_h} = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h}, \quad V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2, \; V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h \sigma_h\right)^2$$

Si se quiere la *afijación de mínima varianza y la expresión de la varianza mínima para el estimador de la proporción y el total de clase* basta sustituir en la fórmula anterior σ_h^2 por $P_h Q_h$.

Afijación óptima

Tenemos:

$$n_h = n \cdot \frac{\dfrac{W_h\sigma_h}{\sqrt{c_h}}}{\displaystyle\sum_{h=1}^{L}\frac{W_h\sigma_h}{\sqrt{c_h}}} = n \cdot \frac{\dfrac{N_h\sigma_h}{\sqrt{c_h}}}{\displaystyle\sum_{h=1}^{L}\frac{N_h\sigma_h}{\sqrt{c_h}}}, \qquad V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h\,\sigma_h/\sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h\sigma_h\sqrt{c_h}\right),$$

$$V(\hat{X}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h\,\sigma_h/\sqrt{c_h}\right)\left(\sum_{h=1}^{L} N_h\sigma_h\sqrt{c_h}\right)$$

Si se quiere la *afijación óptima y la expresión de la varianza mínima para el estimador de la proporción y el total de clase* basta sustituir en las fórmulas anterioriores σ_h^2 por P_hQ_h.

TAMAÑO DE LA MUESTRA PARA MUESTREO SIN REPOSICIÓN

Vamos a analizar ahora el tamaño de muestra estratificada necesario para cometer un determinado error de muestreo conocido de antemano. Distinguiremos los casos de error de muestreo dado con y sin coeficiente de confianza adicional y, además, distinguiremos entre los diferentes tipos de afijación de la muestra.

Tipo de error → *Parámetro* ↓	*Absoluto proporcional*	*Absoluto varianza mínima*	*Absoluto y coeficiente de confianza adicional proporcional*	*Absoluto y coeficiente de confianza adicional varianza mínima*
Media	$\dfrac{\sum_{h=1}^{L} W_hS_h^2}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_hS_h^2}$	$\dfrac{\left(\sum_{h=1}^{L} W_hS_h\right)^2}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_hS_h^2}$	$\dfrac{\sum_{h=1}^{L} W_hS_h^2}{\frac{e^2}{\lambda_\alpha^2} + \frac{1}{N}\sum_{h=1}^{L} W_hS_h^2}$	$\dfrac{\left(\sum_{h=1}^{L} W_hS_h\right)^2}{\frac{e^2}{\lambda_\alpha^2} + \frac{1}{N}\sum_{h=1}^{L} W_hS_h^2}$
Total	$\dfrac{N\sum_{h=1}^{L} N_hS_h^2}{e^2 + \sum_{h=1}^{L} N_hS_h^2}$	$\dfrac{\left(\sum_{h=1}^{L} N_hS_h\right)^2}{e^2 + \sum_{h=1}^{L} N_hS_h^2}$	$\dfrac{N\sum_{h=1}^{L} N_hS_h^2}{\frac{e^2}{\lambda_\alpha^2} + \sum_{h=1}^{L} N_hS_h^2}$	$\dfrac{\left(\sum_{h=1}^{L} N_hS_h\right)^2}{\frac{e^2}{\lambda_\alpha^2} + \sum_{h=1}^{L} N_hS_h^2}$
Proporción	$\dfrac{\sum_{h=1}^{L} W_h\frac{N_h}{N_h-1}P_hQ_h}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h\frac{N_h}{N_h-1}P_hQ_h}$	$\dfrac{\left(\sum_{h=1}^{L} W_h\sqrt{\frac{N_h}{N_h-1}P_hQ_h}\right)^2}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h\frac{N_h}{N_h-1}P_hQ_h}$	$\dfrac{\sum_{h=1}^{L} W_h\frac{N_h}{N_h-1}P_hQ_h}{\frac{e^2}{\lambda_\alpha^2} + \frac{1}{N}\sum_{h=1}^{L} W_h\frac{N_h}{N_h-1}P_hQ_h}$	$\dfrac{\left(\sum_{h=1}^{L} W_h\sqrt{\frac{N_h}{N_h-1}P_hQ_h}\right)^2}{\frac{e^2}{\lambda_\alpha^2} + \frac{1}{N}\sum_{h=1}^{L} W_h\frac{N_h}{N_h-1}P_hQ_h}$
Total de clase	$\dfrac{N\sum_{h=1}^{L} N_h\frac{N_h}{N_h-1}P_hQ_h}{e^2 + \sum_{h=1}^{L} N_h\frac{N_h}{N_h-1}P_hQ_h}$	$\dfrac{\left(\sum_{h=1}^{L} N_h\sqrt{\frac{N_h}{N_h-1}P_hQ_h}\right)^2}{e^2 + \sum_{h=1}^{L} N_h\frac{N_h}{N_h-1}P_hQ_h}$	$\dfrac{N\sum_{h=1}^{L} N_h\frac{N_h}{N_h-1}P_hQ_h}{\frac{e^2}{\lambda_\alpha^2} + \sum_{h=1}^{L} N_h\frac{N_h}{N_h-1}P_hQ_h}$	$\dfrac{\left(\sum_{h=1}^{L} N_h\sqrt{\frac{N_h}{N_h-1}P_hQ_h}\right)^2}{\frac{e^2}{\lambda_\alpha^2} + \sum_{h=1}^{L} N_h\frac{N_h}{N_h-1}P_hQ_h}$

TAMAÑO DE LA MUESTRA PARA MUESTREO CON REPOSICIÓN

Vamos a analizar ahora el tamaño de muestra estratificada con reposición necesario para cometer un determinado error de muestreo conocido de antemano. Distinguiremos los casos de error de muestreo dado con y sin coeficiente de confianza adicional y, además, distinguiremos entre los diferentes tipos de afijación de la muestra.

Tipo de error → *Parámetro* ↓	*Absoluto proporcional*	*Absoluto varianza mínima*	*Absoluto y coeficiente de confianza adicional proporcional*	*Absoluto y coeficiente de confianza adicional varianza mínima*
Media	$\frac{\sum_{h=1}^{L} W_h \sigma_h^2}{e^2}$	$\frac{\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{e^2}$	$\frac{\sum_{h=1}^{L} W_h \sigma_h^2}{e^2 / \lambda_\alpha^2}$	$\frac{\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{e^2 / \lambda_\alpha^2}$
Total	$\frac{N\sum_{h=1}^{L} N_h \sigma_h^2}{e^2}$	$\frac{\left(\sum_{h=1}^{L} N_h \sigma_h\right)^2}{e^2}$	$\frac{N\sum_{h=1}^{L} N_h \sigma_h^2}{e^2 / \lambda_\alpha^2}$	$\frac{\left(\sum_{h=1}^{L} N_h \sigma_h\right)^2}{e^2 / \lambda_\alpha^2}$
Proporción	$\frac{\sum_{h=1}^{L} W_h P_h Q_h}{e^2}$	$\frac{\left(\sum_{h=1}^{L} W_h \sqrt{P_h Q_h}\right)^2}{e^2}$	$\frac{\sum_{h=1}^{L} W_h P_h Q_h}{e^2 / \lambda_\alpha^2}$	$\frac{\left(\sum_{h=1}^{L} W_h \sqrt{P_h Q_h}\right)^2}{e^2 / \lambda_\alpha^2}$
Total de clase	$\frac{N\sum_{h=1}^{L} N_h P_h Q_h}{e^2}$	$\frac{\left(\sum_{h=1}^{L} N_h \sqrt{P_h Q_h}\right)^2}{e^2}$	$\frac{N\sum_{h=1}^{L} N_h P_h Q_h}{e^2 / \lambda_\alpha^2}$	$\frac{\left(\sum_{h=1}^{L} N_h \sqrt{P_h Q_h}\right)^2}{e^2 / \lambda_\alpha^2}$

COMPARACIÓN DE EFICIENCIAS EN MUESTREO ESTRATIFICADO

Muestreo sin reposición

Vamos a realizar ahora comparaciones de eficiencias a partir de la expresión de S^2. Tenemos:

$$S^2 = \sum_{h=1}^{L} W_h S_h^2 + \sum_{h=1}^{L} W_h \left(\overline{X}_h - \overline{X}\right)^2 \Rightarrow \frac{S^2}{n} = \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2 + \frac{1}{n}\sum_{h=1}^{L} W_h \left(\overline{X}_h - \overline{X}\right)^2 \Rightarrow$$

$$\underbrace{(1-f)\frac{S^2}{n}}_{V_{MAS}(\overline{x})} = \underbrace{\frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2}_{V_{MEP}(\overline{x})} + \underbrace{\frac{1-f}{n}\sum_{h=1}^{L} W_h \left(\overline{X}_h - \overline{X}\right)^2}_{\geq 0} \Rightarrow V_{MAS}(\overline{x}) \underset{\substack{\downarrow \\ La\ igualdad\ se\ da \\ si\ \overline{X}_h = \overline{X}\ h=1,\cdots,L}}{\geq} V_{MEP}(\overline{x})$$

Hemos visto que el muestreo estratificado con afijación proporcional es más preciso que el muestreo aleatorio simple, produciéndose la igualdad de precisiones cuando las medias de los estratos son todas iguales. Por tanto, la ganancia en precisión del muestreo estratificado respecto del aleatorio simple será mayor cuanto más distintas entre sí sean las medias de los estratos; es decir, para que el muestreo estratificado sea preciso es conveniente que los estratos sean heterogéneos entre sí en media, afirmación que ya conocíamos desde el comienzo del tema y que constituye una de las especificaciones clásicas en el muestreo estratificado.

$$V_{MEP}(\overline{x}) - V_{MEMV}(\overline{x}) = \underbrace{\frac{1-f}{n}}_{\frac{1}{n}-\frac{1}{N}}\sum_{h=1}^{L} W_h S_h^2 - \left(\frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2\right) =$$

$$\frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h^2 - \left(\sum_{h=1}^{L} W_h S_h\right)^2\right) = \frac{1}{n}\sum_{h=1}^{L} W_h \left(S_h - \overline{S}\right)^2 \underset{\substack{\downarrow \\ La\ igualdad\ se\ da \\ si\ S_h = \overline{S}\ h=1,\cdots,L}}{\geq} 0 \; con \; \overline{S} = \sum_{h=1}^{L} W_h S_h$$

Luego $V_{MEP}(\overline{x}) - V_{MEMV}(\overline{x}) \geq 0 \Rightarrow V_{MEP}(\overline{x}) \geq V_{MEMV}(\overline{x})$

El muestreo estratificado con afijación de mínima varianza es más preciso que el muestreo estratificado con afijación proporcional, produciéndose la igualdad de precisiones cuando las cuasidesviaciones típicas de los estratos son todas iguales. Por tanto, la ganancia en precisión del muestreo estratificado con afijación de mínima varianza respecto del muestreo estratificado con afijación proporcional será mayor cuanto más distintas entre sí sean las cuasidesviaciones típicas de los estratos; es decir, para que el muestreo estratificado sea más preciso es conveniente que los estratos sean heterogéneos entre sí en desviación típica, afirmación que ya conocíamos desde el comienzo del tema y que constituye una de las especificaciones clásicas en el muestreo estratificado.

$$V_{MAS}\ (\bar{x}) \geq V_{MEP}\ (\bar{x}) \geq V_{MEMV}\ (\bar{x})$$

El muestreo estratificado con afijación de mínima varianza es más preciso que el muestreo estratificado con afijación proporcional y que el aleatorio simple, siendo además el estratificado con afijación proporcional más preciso que el aleatorio simple.

$$\underbrace{(1-f)\frac{S^2}{n}}_{V_{MAS}(\bar{x})} = \underbrace{\frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2}_{V_{MEP}(\bar{x})} + \frac{1-f}{n}\sum_{h=1}^{L} W_h \left(\bar{X}_h - \bar{X}\right)^2 =$$

$$V_{MEMV}\ (\bar{x}) + \frac{1}{n}\sum_{h=1}^{L} W_h \left(S_h - \bar{S}\right)^2 + \frac{1-f}{n}\sum_{h=1}^{L} W_h \left(\bar{X}_h - \bar{X}\right)^2$$

El incremento de la eficiencia del muestreo estratificado con afijación de mínima varianza respecto del muestreo aleatorio simple recoge un término debido a la variabilidad de las medias de los estratos y otro debido a la variabilidad de las desviaciones típicas de los estratos. Se produce la igualdad de eficiencias cuando las cuasivarianzas y las medias de los estratos son constantes, y se produce la máxima diferencia de eficiencias cuanto más distintas sean las cuasivarianzas y las medias de los estratos, es decir, cuanto mayor sea la heterogeneidad entre los estratos, tal y como es lógico en muestreo estratificado.

Muestreo con reposición

Vamos a realizar ahora comparaciones de eficiencias a partir de la expresión de σ^2. Tenemos:

$$\sigma^2 = \sum_{h=1}^{L} W_h \sigma_h^2 + \sum_{h=1}^{L} W_h \left(\bar{X}_h - \bar{X}\right)^2 \Rightarrow \underbrace{\frac{\sigma^2}{n}}_{V_{MAS}(\bar{x})} = \underbrace{\frac{1}{n}\sum_{h=1}^{L} W_h \sigma_h^2}_{V_{MEP}(\bar{x})} + \underbrace{\frac{1}{n}\sum_{h=1}^{L} W_h \left(\bar{X}_h - \bar{X}\right)^2}_{\geq 0} \Rightarrow$$

$$V_{MAS}(\bar{x}) \underset{\substack{\text{La igualdad se da} \\ \text{si } \bar{X}_h = \bar{X}\ h=1,\cdots,L}}{\geq} V_{MEP}(\bar{x})$$

Hemos visto que el muestreo estratificado con reposición y afijación proporcional es más preciso que el muestreo aleatorio simple con reposición, produciéndose la igualdad de precisiones cuando las medias de los estratos son todas iguales.

Ahora vamos a comparar la afijación proporcional y de mínima varianza con reposición.

$$V_{MEP}(\bar{x}) - V_{MEMV}(\bar{x}) = \frac{1}{n}\sum_{h=1}^{L} W_h \sigma_h^2 - \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2 = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h^2 - \left(\sum_{h=1}^{L} W_h \sigma_h\right)^2\right)$$

$$= \frac{1}{n}\sum_{h=1}^{L} W_h \left(\sigma_h - \bar{\sigma}\right)^2 \underset{\substack{\text{La igualdad se da} \\ \text{si } S_h = \bar{S}\, h=1,\cdots,L}}{\geq} 0 \; con \bar{\sigma} = \sum_{h=1}^{L} W_h \sigma_h \Rightarrow V_{MEP}(\bar{x}) \geq V_{MEMV}(\bar{x})$$

El muestreo estratificado con reposición y afijación de mínima varianza es más preciso que el muestreo estratificado con reposición y afijación proporcional, produciéndose la igualdad de precisiones cuando las cuasidesviaciones típicas de los estratos son todas iguales.

$$V_{MAS}(\bar{x}) \geq V_{MEP}(\bar{x}) \geq V_{MEMV}(\bar{x})$$

En general el muestreo estratificado con reposición y afijación de mínima varianza es más preciso que el muestreo estratificado con reposición y afijación proporcional y que el aleatorio simple con reposición, siendo además el estratificado con reposición y afijación proporcional más preciso que el aleatorio simple con reposición.

$$\underbrace{\frac{\sigma^2}{n}}_{V_{MAS}(\bar{x})} = \underbrace{\frac{1}{n}\sum_{h=1}^{L} W_h \sigma_h^2}_{V_{MEP}(\bar{x})} + \frac{1}{n}\sum_{h=1}^{L} W_h \left(\bar{X}_h - \bar{X}\right)^2 =$$

$$V_{MEMV}(\bar{x}) + \frac{1}{n}\sum_{h=1}^{L} W_h \left(\sigma_h - \bar{\sigma}\right)^2 + \frac{1}{n}\sum_{h=1}^{L} W_h \left(\bar{X}_h - \bar{X}\right)^2$$

El incremento de la eficiencia del muestreo estratificado con reposición y afijación de mínima varianza respecto del muestreo aleatorio simple con reposición recoge un término debido a la variabilidad de las medias de los estratos y otro debido a la variabilidad de las desviaciones típicas de los estratos. Se produce la igualdad de eficiencias cuando las varianzas y las medias de los estratos son constantes, y se produce la máxima diferencia de eficiencias cuanto más distintas sean las varianzas y las medias de los estratos, es decir, cuanto mayor sea la heterogeneidad entre los estratos, tal y como es lógico en muestreo estratificado.

POSTESTRATIFICACIÓN

Cuando se manejan determinadas variables de estratificación puede ocurrir que no se conozca el estrato a que pertenece una unidad sino hasta después de recoger los datos.

Ejemplos típicos son las características personales como la edad, el sexo, la estatura, etc., y el nivel de educación.

Los tamaños de los estratos N_h se pueden obtener de manera bastante exacta a partir de las estadísticas oficiales, pero las unidades se pueden clasificar en estratos solamente después de conocer los datos de la muestra. Por lo tanto, puede suponerse que los W_h y los N_h son conocidos.

Este método se utiliza cuando se desconocen *a priori* las unidades que pertenecen a cada estrato. Obtenida la muestra, las unidades se asignan al estrato correspondiente. Si los pesos de éstos son conocidos, se puede utilizar el estimador insesgado.

$$\bar{x}' = \sum_{h=1}^{L} W_h \bar{x}_h$$

cuya precisión es similar a la obtenida con la afijación proporcional, siempre que todos los n_h sean grandes; por ejemplo, superiores a 20 unidades. Si de los W_h se conocen sólo las aproximaciones W'_h, el estimador:

$$\bar{x}'' = \sum_{h=1}^{L} W'_h \bar{x}_h$$

será sesgado y la cuantía del sesgo será:

$$E[\bar{x}''] - \bar{X} = \sum_{h=1}^{L} W'_h \bar{X}_h - \sum_{h=1}^{L} W_h \cdot \bar{X}_h = \sum_{h=1}^{L} (W'_h W_h) \cdot \bar{X}_h$$

La acuracidad vendrá dada por el error medio cuadrático

$$E.M.C.(\bar{x}''] = \sum_{h=1}^{L} W'^2_h \cdot \frac{S_h^2}{n_h} \cdot (1 - f_h) + \left[\sum (W'_h - W_h) \bar{X}_h \right]^2$$

El estimador del total es: $\hat{X}'' = \sum_{h=1}^{L} N'_h \bar{x}_h$.

El método de postestratificación puede aplicarse también a una muestra ya estratificada por otro factor, por ejemplo, en cinco regiones geográficas a condición de que los W_h se conozcan separadamente en cada región. Esta estratificación doble se utiliza mucho en las cuentas nacionales de Estados Unidos. Los errores se calculan y estiman mediante:

$$V(\bar{x}'') = \frac{N-n}{N^2 n} \sum_{h=1}^{L} N'_h \cdot S'^2_h + \frac{N-n}{N n^2} \sum_{h=1}^{L} S'^2_h (1 - f'_h)$$

$$V(\hat{X}'') = \frac{N-n}{n} \sum_{h=1}^{L} N'_h \cdot S'^2_h + \frac{N(N-n)}{n^2} \sum_{h=1}^{L} S'^2_h (1 - f'_h)$$

$$\hat{V}(\bar{x}'') = \frac{N-n}{N^2 n} \sum_{h=1}^{L} N'_h \cdot \hat{S}'^2_h + \frac{N-n}{N n^2} \sum_{h=1}^{L} \hat{S}'^2_h (1 - f'_h)$$

$$\hat{V}(\hat{X}'') = \frac{N-n}{n} \sum_{h=1}^{L} N'_h \cdot \hat{S}'^2_h + \frac{N(N-n)}{n^2} \sum_{h=1}^{L} \hat{S}'^2_h (1 - f'_h)$$

Para totales y proporciones cambiamos $\hat{S}_h^2$ por $\frac{n'_h}{n'_h - 1} \hat{P}'_h (1 - \hat{P}'_h)$ y S'^2_h por $\frac{N'_h}{N'_h - 1} P'_h (1 - P'_h)$. El apóstrofe indica siempre valor de postestratificación.

PROBLEMAS RESUELTOS

4.1. Una empresa publicitaria está interesada en medir la influencia de la publicidad televisiva en un municipio y decide realizar una encuesta por muestreo para estimar el número promedio de horas por semana que se ve la televisión en los hogares del municipio. Éste comprende dos pueblos A y B y un área rural, y se sabe que existen 155 hogares en el pueblo A, 62 en el pueblo B y 93 en el área rural. La empresa publicitaria tiene tiempo y dinero suficientes para entrevistar 30 hogares (20 del pueblo A, 8 del pueblo B y 12 del área rural) midiendo en cada uno el tiempo que se ve la televisión en horas por semana. Se obtienen los datos siguientes:

Puebla A (estrato I)→ 35 28 26 41 43 29 32 37 36 25 29 31 39 38 40 45 28 27 35 34
Pueblo B (estrato II)→ 27 4 49 10 15 41 25 30
Área rural (estrato III)→ 8 15 21 7 14 30 20 11 12 32 34 24

Estimar el tiempo promedio que se ve la televisión, en horas por semana, en cada uno de los estratos y en todo el municipio fijando límites para el error de estimación a través de intervalos de confianza al 95%.

Comenzamos introduciendo los datos como tres columnas, una por cada estrato, en una hoja de cálculo de Excel. A continuación, para calcular los estadísticos necesarios en cada estrato, en el menú *Herramientas* de Excel elegimos *Análisis de datos*, seleccionamos *Estadística descriptiva* y rellenamos la pantalla de entrada como se indica en la Figura 13-1. Al pulsar *Aceptar* se obtienen los estadísticos muestrales por estrato de la Figura 13-2. Se observa que el tiempo promedio que se ve la televisión en el pueblo A es 33,9 horas por semana, en el pueblo es 20,33 y en la zona rural es 19. Las cuasivarianzas muestrales son 33,3578, 285 y 87,63 horas por semana, respectivamente, en cada estrato, y al dividirlas por el tamaño muestral seleccionado en cada estrato obtenemos los errores de los estimadores en cada estrato suponiendo muestreo con reposición (33,35/20 = 1,667, 285/8 = 35,62 y 87,63/12 = 7,3). Como los coeficientes de asimetría y curtosis en cada estrato están en el intervalo [–2,2], puede suponerse normalidad, con lo que los límites para el error de estimación en cada estrato (suponiendo muestreo con reposición) serán los radios de los intervalos de confianza al 95%, es decir, 2,7829, 12,97 y 5,94, respectivamente. Si el muestreo es sin reposición, las varianzas en cada estrato hay que multiplicarlas por $(1 - n_h/N_h)$ $h = 1, 2, 3$.

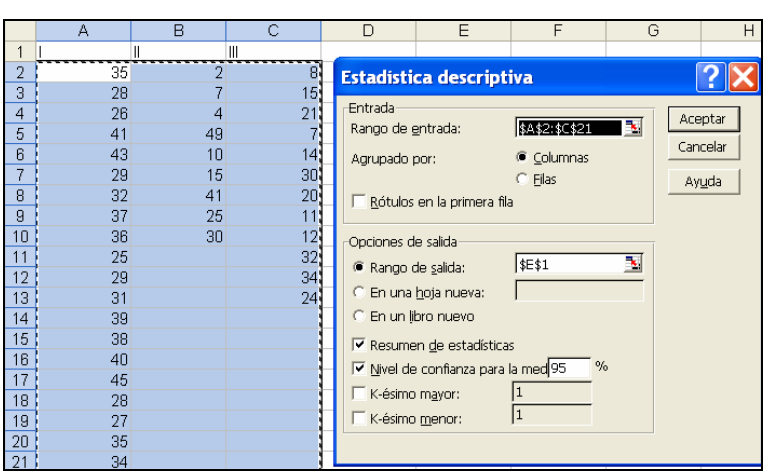


Figura 13-1

	E	F	G	H	I	J
1	*Columna1*		*Columna2*		*Columna3*	
2						
3	**Media**	**33,9**	**Media**	**20,3333333**	**Media**	**19**
4	Error típico	1,32962203	Error típico	5,62731434	Error típico	2,70241194
5	Mediana	34,5	Mediana	15	Mediana	17,5
6	Moda	35	Moda	#N/A	Moda	#N/A
7	Desviación estándar	5,94625048	Desviación estándar	16,881943	Desviación estándar	9,36142957
8	**Varianza de la muestra**	**35,3578947**	**Varianza de la muestra**	**285**	**Varianza de la muestra**	**87,6363636**
9	**Curtosis**	**-1,0468008**	**Curtosis**	**-0,9607396**	**Curtosis**	**-1,2178641**
10	**Coeficiente de asimetría**	**0,20737524**	**Coeficiente de asimetría**	**0,64283214**	**Coeficiente de asimetría**	**0,3901415**
11	Rango	20	Rango	47	Rango	27
12	Mínimo	25	Mínimo	2	Mínimo	7
13	Máximo	45	Máximo	49	Máximo	34
14	Suma	678	Suma	183	Suma	228
15	Cuenta	20	Cuenta	9	Cuenta	12
16	**Nivel de confianza(95,0%)**	**2,78293175**	**Nivel de confianza(95,0%)**	**12,9766185**	**Nivel de confianza(95,0%)**	**5,94797159**
17						

Figura 13-2

Para hallar la estimación del tiempo promedio que se ve la televisión en todo el municipio en horas por semana y su error para muestreo sin reposición, se tendrán en cuenta las siguientes expresiones:

$$\hat{\bar{X}}_{st} = \bar{x}_{st} = \sum_{h=1}^{L} \underbrace{\frac{N_h}{N}}_{W_h} \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi} = \sum_{h=1}^{L} W_h \bar{x}_h$$

$$\hat{V}\left(\bar{X}_{st}\right) = \sum_{h=1}^{L} W_h^2 \cdot \left(1 - f_h\right) \cdot \frac{\hat{S}_h^2}{n_h}$$

La Figura 13-3 presenta las fórmulas para el cálculo del estimador de la media estratificada para todo el municipio, su error de muestreo y el radio del intervalo de confianza al 95%. La Figura 13-4 presenta los resultados.

	L	M	N	O	P	Q	R	S
1	Nh	*nh*	*Wh*	*Wh(1-nh/Nh)*	*Sh2*	*Wh2(1-nh/Nh)Sh2/nh*	*mediah*	*Confianza*
2	155	20	=L2/$L$5	=N2*(1-M2/L2)	=$F$8	=N2*O2*P2/M2	=$F$3	
3	62	8	=L3/$L$5	=N3*(1-M3/L3)	=$H$8	=N3*O3*P3/M3	=$H$3	
4	93	12	=L4/$L$5	=N4*(1-M4/L4)	=$J$8	=N4*O4*P4/M4	=$J$3	
5	=SUMA(Nh)					**=SUMA(Q2:Q4)**	**=PROMEDIO(mediah)**	**=2*RAIZ(Q5)**

Figura 13-3

	L	M	N	O	P	Q	R	S
1	Nh	*nh*	*Wh*	*Wh(1-nh/Nh)*	*Sh2*	*Wh2(1-nh/Nh)Sh2/nh*	*mediah*	*Confianza*
2	155	20	0,5	0,435483871	35,3578947	0,384944822	33,9	
3	62	8	0,2	0,174193548	285	1,241129032	20,3333333	
4	93	12	0,3	0,261290323	87,6363636	0,572463343	19	
5	310					**2,198537197**	**24,4111111**	**2,96549301**

Figura 13-4

La estimación del tiempo promedio que se ve la televisión en todo el municipio en horas por semana en muestreo con reposición es la misma que sin reposición y su error de muestreo se calcula mediante la siguiente expresión:

$$\hat{V}\left(\bar{X}_{st}\right) = \sum_{h=1}^{L} W_h^2 \cdot \frac{\hat{S}_h^2}{n_h}$$

La Figura 13-5 presenta las fórmulas y la Figura 13-6 presenta los resultados.

	T	U
1	*Wh2Sh2/nh*	*confianza*
2	=N2^2*P2/M2	
3	=N3^2*P3/M3	
4	=N4^2*P4/M4	
5	**=SUMA(T2:T4)**	**=2*RAIZ(T5)**

Figura 13-5

	T	U
1	*Wh2Sh2/nh*	*confianza*
2	0,441973684	
3	1,425	
4	0,657272727	
5	**2,524246411**	**3,177575435**

Figura 13-6

4.2. Consideramos los salarios anuales (variable X) en miles de euros de 500 trabajadores de una empresa se obtiene la siguiente distribución de frecuencias:

X_i	n_i
2	100
3	80
5	200
10	30
20	30
50	30
100	20
200	10

Se estratifica la población en grupos homogéneos de ganancias salariales utilizando como variable de estratificación el propio salario anual mediante el criterio dado por $2 \leq X < 10$, $10 \leq X < 100$, $100 \leq X \leq 200$. Realizar las afijaciones de mínima varianza sin y con reposición de una muestra de tamaño 100 cuando se estima el salario anual medio. Analizar las precisiones y justificar los resultados.

Comenzamos realizando los cálculos por estratos necesarios para la resolución del problema. Tenemos:

Estratos ↓	S_h	S_h^2	σ_h	σ_h^2	N_h
I	1,32	1,75	1,32	1,74	380
II	17,1	292,13	16,99	288,88	90
III	47,95	2298,85	47,14	2222,22	30

Afijación de mínima varianza sin reposición

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \Rightarrow \begin{cases} n_1 = 100 \cdot \dfrac{N_1 S_1}{N_1 S_1 + N_2 S_2 + N_3 S_3} \cong 15 \\ n_2 = 100 \cdot \dfrac{N_2 S_2}{N_1 S_1 + N_2 S_2 + N_3 S_3} \cong 44 \\ n_3 = 100 \cdot \dfrac{N_3 S_3}{N_1 S_1 + N_2 S_2 + N_3 S_3} \cong 41 \end{cases}$$

Se observa que el número de unidades a seleccionar para la muestra en el tercer estrato es superior al número de unidades de dicho estrato.

Ante esta circunstancia seleccionamos para la muestra las 30 unidades del tercer estrato; es decir, todas las unidades del tercer estrato van a ser autorrepresentadas. Pero ahora las 70 unidades restantes de la muestra han de repartirse mediante afijación de mínima varianza entre los dos primeros estratos. Tendremos:

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \Rightarrow \begin{cases} n_1 = 70 \cdot \dfrac{N_1 S_1}{N_1 S_1 + N_2 S_2} \cong 17 \\ n_2 = 70 \cdot \dfrac{N_2 S_2}{N_1 S_1 + N_2 S_2} \cong 53 \end{cases}$$

Por tanto, la nueva afijación es $n_1 = 17$, $n_2 = 53$ y $n_3 = 30$. Para hallar la varianza del estimador de la media para esta afijación sin reposición hemos de tener en cuenta que los estratos con sus unidades autorrepresentadas no intervienen en el cálculo de las varianzas. Como el tercer estrato no interviene en el valor de la varianza, calculamos $W'_1 = \frac{N_1}{N'} = \frac{380}{470} = 0{,}8085$ y $W'_2 = \frac{N_2}{N'} = \frac{90}{470} = 0{,}1915$. La varianza será:

$$V\left(\hat{\bar{X}}\right) = \frac{1}{n'}\left(\sum_{h=1}^{2} W'_h S_h\right)^2 - \frac{1}{N'}\sum_{h=1}^{2} W'_h S_h^2 = 0{,}184064 .$$

Afijación de mínima varianza con reposición

Realizaremos la afijación de mínima varianza con reposición como sigue:

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h} \Rightarrow \begin{cases} n_1 = 100 \cdot \dfrac{N_1 \sigma_1}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} \cong 15 \\ n_2 = 100 \cdot \dfrac{N_2 \sigma_2}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} \cong 44 \\ n_3 = 100 \cdot \dfrac{N_3 \sigma_3}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} \cong 41 \end{cases}$$

Se observa que la afijación coincide exactamente con la obtenida para muestreo sin reposición. Ahora el número de unidades a seleccionar para la muestra en el tercer estrato vuelve a ser superior al número de unidades de dicho estrato, pero como el muestreo es con reposición, es posible seguir haciendo extracciones porque las unidades se reponen a la población cuando se extrae y nunca se acabarán. El valor de la varianza mínima será ahora:

$$V\left(\hat{\bar{X}}\right) = \frac{1}{n}\left(\sum_{h=1}^{3} W_h \sigma_h\right)^2 = \frac{1}{100}\left(\frac{380}{500} \cdot 1{,}32 + \frac{90}{500} \cdot 16{,}99 + \frac{30}{500} 47{,}14\right)^2 = 0{,}47469344$$

No obstante, si se exige que las unidades seleccionadas sean distintas, seleccionamos para la muestra las 30 unidades del tercer estrato; es decir, todas las unidades del tercer estrato van a ser autorrepresentadas. Pero ahora las 70 unidades restantes de la muestra han de repartirse mediante afijación de mínima varianza con reposición entre los dos primeros estratos. Tendremos:

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h} \Rightarrow \begin{cases} n_1 = 70 \cdot \dfrac{N_1 \sigma_1}{N_1 \sigma_1 + N_2 \sigma_2} \cong 17 \\ n_2 = 70 \cdot \dfrac{N_2 \sigma_2}{N_1 \sigma_1 + N_2 \sigma_2} \cong 53 \end{cases}$$

Por tanto, la nueva afijación es $n_1 = 17$, $n_2 = 53$ y n_3=30. Para hallar la varianza del estimador de la media para esta afijación con reposición hemos de tener en cuenta que los estratos con sus unidades autorrepresentadas no intervienen en el cálculo de las varianzas. Como el tercer estrato no interviene en el valor de la varianza calculamos $W'_1 = \frac{N_1}{N'} = \frac{380}{470} = 0{,}8085$ y $W'_2 = \frac{N_2}{N'} = \frac{90}{470} = 0{,}1915$. La varianza será:

$$V\left(\hat{\bar{X}}\right) = \frac{1}{n'}\left(\sum_{h=1}^{2} W'_h \sigma_h\right)^2 = \frac{1}{70}(0{,}8085 \cdot 1{,}32 + 0{,}1915 \cdot 16{,}99)^2 = 0{,}266705.$$

Las afijaciones coinciden para muestreo con y sin reposición, pero el muestreo sin reposición resulta más preciso, ya que tiene menor varianza (tanto si se exigen unidades distintas, como en caso contrario).

Se observa que, aunque haya estratos con todas sus unidades autorrepresentadas, el muestreo sin reposición sigue siendo más preciso que el muestreo con reposición.

4.3. Las mil unidades de una población se clasifican en tres estratos para los que se conocen los datos de la tabla adjunta:

Estratos ↓	σ_i	W_i
I	4	0,6
II	12	0,3
III	80	0,1

Se pide:

1) Determinar el tamaño de muestra que con afijación proporcional proporciona una varianza del estimador de la media igual a 5, considerando muestreo con y sin reposición. Realizar las respectivas afijaciones proporcionales. ¿Qué resultados se obtendrían con afijación de mínima varianza? Realizar las respectivas afijaciones de mínima varianza. Comentar todos los resultados y compararlos.

2) Determinar el tamaño de muestra para afijación óptima con costes C_1=1000, C_2=1200 y C_3=2000, considerando el muestreo con y sin reposición. Realizar las respectivas afijaciones óptimas. Comprobar que los resultados coinciden para costes unitarios con los de afijación de mínima varianza.

Como es habitual en los problemas de muestreo estratificado, comenzamos recopilando los datos necesarios para el problema.

$W_1=0,6=N_1/N\Rightarrow N_1=600$ $\quad \sigma_1^2=16=(N_1-1)S_1^2/N_1\Rightarrow S_1^2=6,02\Rightarrow S_1=4,003$
$W_2=0,3=N_2/N\Rightarrow N_2=300$ $\quad \sigma_2^2=144=(N_2-1)S_2^2/N_2\Rightarrow S_2^2=144,5\Rightarrow S_2=12,02$
$W_3=0,1=N_3/N\Rightarrow N_3=100$ $\quad \sigma_3^2=6400=(N_3-1)S_3^2/N_3\Rightarrow S_3^2=6464,6\Rightarrow S_3=80,4$

Tenemos entonces:

Estratos ↓	S_h	S_h^2	σ_h	σ_h^2	N_h	W_h
I	4,003	6,02	4	16	600	0,6
II	12,02	144,5	12	144	300	0,3
III	80,4	6464,6	80	6400	100	0,1

Afijación proporcional sin reposición

$$e^2=V\left(\hat{\bar{X}}\right)=\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{h=1}^{L}W_hS_h^2\Rightarrow n=\frac{\sum_{h=1}^{L}W_hS_h^2}{e^2+\frac{1}{N}\sum_{h=1}^{L}W_hS_h^2}\cong 122$$

Una vez hallado el tamaño de muestra, realizamos la afijación como sigue:

$$n_h=kN_h \text{ con } k=\frac{n}{N}=\frac{122}{1000}=0,122\Rightarrow\begin{cases}n_1=kN_1=0,122\cdot 600\cong 73\\ n_2=kN_2=0,122\cdot 300\cong 37\\ n_3=kN_3=0,122\cdot 100\cong 12\end{cases}$$

Afijación proporcional con reposición

$$e^2=V\left(\hat{\bar{X}}\right)=\frac{1}{n}\sum_{h=1}^{L}W_h\sigma_h^2\Rightarrow n=\frac{\sum_{h=1}^{L}W_h\sigma_h^2}{e^2}\cong 139$$

Se observa que el tamaño muestral necesario para cometer el mismo error que sin reposición es ahora superior. Ello es debido a que el muestreo con reposición es menos preciso que el muestreo sin reposición. Una vez hallado el tamaño de muestra realizamos la afijación proporcional como sigue:

$$n_h=kN_h \text{ con } k=\frac{n}{N}=\frac{139}{1000}=0,139\Rightarrow\begin{cases}n_1=kN_1=0,139\cdot 600\cong 83\\ n_2=kN_2=0,139\cdot 300\cong 42\\ n_3=kN_3=0,139\cdot 100\cong 14\end{cases}$$

Afijación de mínima varianza sin reposición

$$e^2=V\left(\hat{\bar{X}}\right)=\frac{1}{n}\left(\sum_{h=1}^{L}W_hS_h\right)^2-\frac{1}{N}\sum_{h=1}^{L}W_hS_h^2\Rightarrow n=\frac{\left(\sum_{h=1}^{L}W_hS_h\right)^2}{e^2+\frac{1}{N}\sum_{h=1}^{L}W_hS_h^2}=35$$

Una vez hallado el tamaño de muestra, realizamos la afijación de mínima varianza como sigue:

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \Rightarrow \begin{cases} n_1 = 35 \cdot \dfrac{N_1 S_1}{N_1 S_1 + N_2 S_2 + N_3 S_3} \cong 6 \\ n_2 = 35 \cdot \dfrac{N_2 S_2}{N_1 S_1 + N_2 S_2 + N_3 S_3} \cong 9 \\ n_3 = 35 \cdot \dfrac{N_3 S_3}{N_1 S_1 + N_2 S_2 + N_3 S_3} \cong 20 \end{cases}$$

Afijación de mínima varianza con reposición

$$e^2 = V\left(\hat{\bar{X}}\right) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{e^2} \cong 40$$

Se observa que el tamaño muestral necesario para cometer el mismo error que sin reposición es ahora superior. Una vez hallado el tamaño de muestra realizamos la afijación de mínima varianza como sigue:

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h} \Rightarrow \begin{cases} n_1 = 35 \cdot \dfrac{N_1 \sigma_1}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} \cong 7 \\ n_2 = 35 \cdot \dfrac{N_2 \sigma_2}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} \cong 10 \\ n_3 = 35 \cdot \dfrac{N_3 \sigma_3}{N_1 \sigma_1 + N_2 \sigma_2 + N_3 \sigma_3} \cong 23 \end{cases}$$

Afijación óptima sin reposición

$$V(\bar{x}_{st}) = e^2 = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right) - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)}{e^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2} \cong 35$$

Una vez hallado el tamaño de muestra, realizamos la afijación óptima como sigue:

$$n_h = n \cdot \frac{N_h S_h / \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{C_h}} \Rightarrow \begin{cases} n_1 = 35 \cdot \dfrac{N_1 S_1 / \sqrt{C_1}}{N_1 S_1 / \sqrt{C_1} + N_2 S_2 / \sqrt{C_2} + N_3 S_3 / \sqrt{C_3}} \cong 7 \\ n_2 = 35 \cdot \dfrac{N_2 S_2}{N_1 S_1 / \sqrt{C_1} + N_2 S_2 / \sqrt{C_2} + N_3 S_3 / \sqrt{C_3}} \cong 10 \\ n_3 = 35 \cdot \dfrac{N_3 S_3}{N_1 S_1 / \sqrt{C_1} + N_2 S_2 / \sqrt{C_2} + N_3 S_3 / \sqrt{C_3}} \cong 18 \end{cases}$$

Afijación óptima con reposición

$$V(\bar{x}_{st}) = e^2 = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h \sigma_h \sqrt{c_h}\right) \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h \sigma_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h \sigma_h \sqrt{c_h}\right)}{e^2} = 40$$

Se observa que el tamaño muestral necesario para cometer el mismo error que sin reposición es ahora superior. Una vez hallado el tamaño de muestra realizamos la afijación óptima como sigue:

$$n_h = n \cdot \frac{N_h \sigma_h / \sqrt{C_h}}{\sum_{h=1}^{L} N_h \sigma_h / \sqrt{C_h}} \Rightarrow \begin{cases} n_1 = 40 \cdot \dfrac{N_1 \sigma_1 / \sqrt{C_1}}{N_1 \sigma_1 / \sqrt{C_1} + N_2 \sigma_2 / \sqrt{C_2} + N_3 \sigma_3 / \sqrt{C_3}} \cong 8 \\ n_2 = 40 \cdot \dfrac{N_2 \sigma_2}{N_1 \sigma_1 / \sqrt{C_1} + N_2 \sigma_2 / \sqrt{C_2} + N_3 \sigma_3 / \sqrt{C_3}} \cong 12 \\ n_3 = 40 \cdot \dfrac{N_3 \sigma_3}{N_1 \sigma_1 / \sqrt{C_1} + N_2 \sigma_2 / \sqrt{C_2} + N_3 \sigma_3 / \sqrt{C_3}} \cong 20 \end{cases}$$

Si utilizamos costes unitarios los cálculos son exactamente los mismos que para la afijación de mínima varianza, luego los resultados también lo son. Se observa que tanto en muestreo con reposición como sin reposición la afijación que menos tamaño muestral necesita para cometer un determinado error de muestreo es la afijación de mínima varianza, y en este caso también la óptima.

4.4. Una empresa de publicidad quiere estimar la proporción de hogares en un municipio donde se ve cierto programa televisivo. El municipio tiene en total 310 hogares y es dividido en tres estratos. Se selecciona una muestra estratificada de $n = 40$ hogares con afijación proporcional. Estimar la proporción de hogares en el municipio donde se ve el programa televisivo estimando los errores absoluto y relativo cometidos. Datos:

Estratos	*Tamaños muestrales*	*Nº de hogares donde se ve el programa*	$\hat{P}_h$
1	$n_1 = 20$	16	0,80
2	$n_2 = 8$	2	0,25
3	$n_3 = 12$	6	0,50

Como la selección de la muestra se realiza con afijación proporcional, se tiene:

$$n_h = kN_h \text{ con } k = \frac{n}{N} = \frac{40}{310} = 0{,}129 \Rightarrow \begin{cases} N_1 = \dfrac{n_1}{k} = \dfrac{20}{0{,}129} \cong 155 \\ N_2 = \dfrac{n_2}{k} = \dfrac{8}{0{,}129} \cong 62 \\ N_1 = \dfrac{n_3}{k} = \dfrac{12}{0{,}129} \cong 93 \end{cases}$$

Ya podemos estimar la proporción de hogares en el municipio donde se ve el programa televisivo de la siguiente forma:

$$\hat{P}_{st} = \sum_{h=1}^{3} W_h \hat{P}_h = \sum_{h=1}^{3} \frac{N_h}{N} \hat{P}_h = \frac{155}{310} 0{,}80 + \frac{62}{310} 0{,}25 + \frac{93}{310} 0{,}50 = 0{,}60 \quad (60\%)$$

Resulta que en el 60% de los hogares del municipio se ve el programa televisivo.

Para calcular el error absoluto de esta estimación hallamos la estimación de la varianza del estimador de la proporción. Se tiene:

$$\hat{V}(\hat{P}_{st}) = \sum_{h=1}^{3} \frac{N_h^2}{N^2} \hat{V}(\hat{P}_h) = \sum_{h=1}^{3} \frac{N_h^2}{N^2} (1 - \frac{n_h}{N_h}) \frac{\hat{P}_h (1 - \hat{P}_h)}{n_h - 1} = 0{,}0045$$

El error relativo sería $\hat{C}v(\hat{P}_{st}) = \dfrac{\sqrt{\hat{V}(\hat{P}_{st})}}{\hat{P}_{st}} \cdot 100 = \dfrac{\sqrt{0{,}0045}}{0{,}60} \cdot 100 = 11{,}18\%$.

4.5. Una empresa publicitaria está interesada en determinar lo que debe enfatizar la publicidad televisiva en un determinado municipio, y decide realizar una encuesta por muestreo para estimar el número promedio de horas por semana que se ve la televisión en los hogares del municipio. Éste comprende dos pueblos, pueblo A y pueblo B, y un área rural. El pueblo A circunda una fábrica, y la mayoría de los hogares son de trabajadores fabriles con niños en edad escolar. El pueblo B es un suburbio exclusivo de una ciudad vecina y consta de habitantes más viejos con pocos niños en casa. Existen 155 hogares en el pueblo A, 62 en el pueblo B y 93 en el área rural. Se pide:

1. Analizar los méritos de usar muestreo aleatorio estratificado en esa situación.

2. Supóngase que se lleva a cabo la encuesta planificada. La empresa publicitaria tiene tiempo y dinero suficientes para entrevistar $n = 40$ hogares, y decide seleccionar muestras aleatorias de tamaño $n_1 = 20$ del pueblo A, $n_2 = 8$ del pueblo B, y $n_3 = 12$ del área rural. Se seleccionan las muestras irrestrictas aleatorias y se realizan las entrevistas. Los resultados, con mediciones del tiempo que se ve la televisión en horas por semana, son los siguientes:

Estrato 1(pueblo A)	35	43	36	39	28	28	29	25	38	27	26	32	29	40	35	41	37	31	45
Estrato 2 (pueblo B)	27	15	4	41	49	25	10	30											
Estrato 3 (pueblo C)	8	14	12	15	30	32	21	20	34	7	11	24							

Estimar el tiempo promedio que se ve televisión, en horas por semana, para (a) todos los hogares del municipio y (b) todos los hogares en el pueblo B. En ambos casos fijar un límite para el error de estimación.

3. Estimar el número total de horas por semana que las familias del municipio dedican a ver la televisión fijando un límite para el error de estimación.

Comenzamos recopilando la información necesaria para el problema en la tabla siguiente:

Estratos ↓	n_h	S_h^2	S_h	$\bar{x}_h$	N_h
1	20	35,358	5,946	33,9	155
2	8	232,411	15,245	25,125	62
3	12	87,636	9,361	19	93

En cuanto al primer apartado, podemos decir que la población de hogares se ubica en tres grupos naturales, dos pueblos y un área rural, de acuerdo con su localización geográfica. Por lo tanto, la población dividida en tres estratos es bastante natural, lo que lleva a que los elementos de cada estrato deben de presentar tendencias similares de comportamiento entre ellos mismos (homogeneidad dentro). Se espera relativamente poca variabilidad en el número de horas que se ve la televisión en los hogares de cada grupo, lo que hace aplicable el muestro estratificado. Por otro lado, los estratos son adecuados por conveniencia administrativa para seleccionar las muestras y para ejecutar el trabajo de campo. Además, la empresa publicitaria puede obtener estimaciones por separado del número promedio de horas que se ve la televisión en cada estrato.

Para estimar el promedio de horas por semana que se ve la televisión en todo el municipio, utilizamos el estimador de la media estratificada:

$$\bar{x}_{st} = \sum_{h=1}^{3} W_h \bar{x}_h = \frac{155}{310} 33{,}9 + \frac{62}{310} 25{,}125 + \frac{93}{310} 19 = 27{,}7$$

El error de esta estimación será:

$$\hat{V}(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h} = \left(\frac{155}{310}\right)^2 \left(1 - \frac{20}{155}\right) \frac{35{,}3}{20} + \left(\frac{62}{310}\right)^2 \left(1 - \frac{8}{62}\right) \frac{232{,}4}{8} + \left(\frac{93}{310}\right)^2 \left(1 - \frac{12}{93}\right) \frac{87{,}6}{12} = 1{,}97$$

Un intervalo de confianza al 95% ($\lambda_\alpha \cong 2$) será el siguiente:

$$\bar{x}_{st} \pm \lambda_\alpha \sqrt{\hat{V}(\bar{x}_{st})} = 27{,}7 \pm 2 \sqrt{1{,}97} = 27{,}7 \pm 2{,}8$$

Por lo tanto, estimamos que el número promedio de horas que se ve la televisión en los hogares del municipio es de 27,7 horas, con un error de muestreo de $\sqrt{1{,}97} = 1{,}4$ horas y un límite para el error de estimación de ±2,8 horas.

Las ocho observaciones del estrato relativo al pueblo B forman una muestra aleatoria simple para la que podemos aplicar las fórmulas del muestreo irrestricto aleatorio. Tenemos:

$$\bar{x}_2 = 25{,}125$$

$$\hat{V}(\bar{x}_2) = (1 - f_2) \frac{\hat{S}_2^2}{n_2} = \left(1 - \frac{8}{62}\right) \frac{232{,}4}{8} = 25{,}5$$

$$\bar{x}_2 \pm \lambda_\alpha \sqrt{\hat{V}(\bar{x}_2)} = 25{,}125 \pm 10{,}1$$

Por lo tanto, estimamos que el número promedio de horas que se ve la televisión en el pueblo B es de 25,5 horas, con un error de muestreo de $\sqrt{25{,}5} = 5{,}05$ horas y un límite para el error de estimación de ±10,1 horas. El límite del error de estimación es más grande en este caso porque la variabilidad del estrato es grande y su tamaño es pequeño. Se observa que la estimación en todo el municipio es buena, pero en el estrato 2 es peor.

El número total de horas estimado que las familias del municipio dedican a ver la televisión será:

$$\hat{X}_{st} = N\bar{x}_{st} = 300(27{,}7) = 8587 \text{ horas}$$

El error de esta estimación será:

$$V\left(\hat{X}_{st}\right) = N^2\hat{V}\left(\bar{x}_{st}\right) = 300^2\left(1{,}97\right) = 189278{,}56$$

Un intervalo de confianza al 95% ($\lambda_\alpha \cong 2$) será el siguiente:

$$\hat{X}_{st} \pm \lambda_\alpha\sqrt{\hat{V}\left(\hat{X}_{st}\right)} = 8587 \pm 2\sqrt{189278{,}56} = 8587 \pm 870$$

Por lo tanto, estimamos que el número total de horas que se ve la televisión en los hogares del municipio es de 8587 horas, con un error de muestreo de $\sqrt{189278{,}56} = 435$ horas y un límite para el error de estimación de ±870 horas.

Cuando se estiman totales es conveniente relativizar los errores, para que sean más comprensibles. En nuestro caso, el error relativo será:

$$\hat{C}v\left(\hat{X}_{st}\right) = \frac{\sqrt{\hat{V}\left(\hat{X}_{st}\right)}}{\hat{X}_{st}}100 = \frac{\sqrt{189278{,}56}}{8587}100 = 5\%$$

Se trata de un error muy aceptable.

4.6. La empresa publicitaria del ejercicio anterior comprobó que cuesta más obtener una observación del área rural que una del pueblo A o del B. El incremento es debido a los costos de traslado de un hogar rural a otro. El costo por observación en cada pueblo se ha estimado en 9 euros (esto es, $c_1 = c_2 = 9$), y los costos por observación en el área rural se han estimado en 16 euros (esto es, $c_3 = 16$). Las desviaciones estándar por estrato (aproximadas por las varianzas muestrales de una encuesta previa) son $\sigma_1 \approx 5$, $\sigma_2 \approx 15$ y $\sigma_3 \approx 10$. Halle el tamaño de muestra total n y los tamaños de muestra para los estratos n_1, n_2 y n_3, que permiten a la empresa estimar, al mínimo costo, el tiempo promedio que se ve televisión, con un límite para el error de estimación igual a 2 horas.

Supongamos que la firma publicitaria decide utilizar entrevistas por teléfono en lugar de entrevistas personales, porque todos los hogares en el municipio tienen teléfono y este método reduce los costos. El costo de obtener una observación es entonces el mismo en los tres estratos y la empresa desea estimar en este caso la media poblacional μ con un límite para el error de estimación igual a 2 horas. Encuentre el tamaño aproximado de la muestra n y los tamaños de muestra para los estratos n_1, n_2 y n_3.

Supongamos ahora que la empresa publicitaria considera que las varianzas aproximadas que se usaron en los ejemplos previos son erróneas y que las varianzas de los estratos son iguales. El valor común de σ_i fue aproximado por 10 en un estudio preliminar. Se van a efectuar entrevistas por teléfono, por lo que los costos serán iguales en todos los estratos. La empresa desea estimar el número promedio de horas por semana que se ve la televisión en los hogares del municipio, con un límite para el error de estimación igual a 2 horas. Determine el tamaño de muestra y los tamaños de estratos necesarios para lograr esta exactitud.

En primer lugar observamos que, como el límite del error de estimación es 2, tenemos:

$$2\sqrt{\hat{V}(\bar{x}_{st})} = 2 \Rightarrow \hat{V}(\bar{x}_{st}) = 1$$

Como estamos en afijación óptima sin reposición, el tamaño de muestra necesario para cometer un error de muestreo unitario para estimar la media vendrá dado por:

$$V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right) - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

Aproximando las cuasivarianzas por las desviaciones estándar por estrato tenemos (los W_h son los del problema anterior):

$$n = \frac{\left(\sum_{h=1}^{L} W_h \sigma_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h \sigma_h \sqrt{c_h}\right)}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h \sigma_h^2} = 57{,}42 \approx 58$$

Ahora realizamos la afijación óptima como sigue:

$$n_h = n \cdot \frac{N_h \sigma_h / \sqrt{C_h}}{\sum_{h=1}^{L} N_h \sigma_h / \sqrt{C_h}} \Rightarrow \begin{cases} n_1 = 58 \cdot \dfrac{N_1 \sigma_1 / \sqrt{C_1}}{N_1 \sigma_1 / \sqrt{C_1} + N_2 \sigma_2 / \sqrt{C_2} + N_3 \sigma_3 / \sqrt{C_3}} \cong 18 \\ n_2 = 58 \cdot \dfrac{N_2 \sigma_2}{N_1 \sigma_1 / \sqrt{C_1} + N_2 \sigma_2 / \sqrt{C_2} + N_3 \sigma_3 / \sqrt{C_3}} \cong 23 \\ n_3 = 58 \cdot \dfrac{N_3 \sigma_3}{N_1 \sigma_1 / \sqrt{C_1} + N_2 \sigma_2 / \sqrt{C_2} + N_3 \sigma_3 / \sqrt{C_3}} \cong 17 \end{cases}$$

En caso de utilizar entrevista telefónica, los costes unitarios por estrato son iguales, con lo que la afijación óptima coincide con la afijación de mínima varianza. En este caso, el tamaño de muestra para cometer un error de muestreo unitario será:

$$e^2 = V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

Aproximando las cuasivarianzas por las desviaciones estándar por estrato tenemos (los W_h son los del problema anterior):

$$n = \frac{\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h \sigma_h^2} = 56{,}34 \approx 57$$

Una vez hallado el tamaño de muestra, realizamos la afijación de mínima varianza como sigue:

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h} \Rightarrow \begin{cases} n_1 = 57 \cdot \dfrac{N_1 \sigma_1}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3} \cong 17 \\ n_2 = 57 \cdot \dfrac{N_2 \sigma_2}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3} \cong 20 \\ n_3 = 57 \cdot \dfrac{N_3 S_3}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3} \cong 20 \end{cases}$$

Aunque el tamaño de muestra sigue siendo muy parecido al del apartado anterior, la afijación cambia, tomándose más observaciones del área rural ya que ahora no tienen un coste más alto.

Si, además de utilizar costes unitarios, suponemos que la variabilidad en los estratos es unitaria, podemos aproximar la afijación óptima y la de mínima varianza por la proporcional, ya que en este caso coinciden las tres. Entonces, el tamaño de muestra para cometer un error de muestreo unitario será:

$$e^2 = V(\bar{x}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\sum_{h=1}^{L} W_h S_h^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

Aproximando las cuasivarianzas por las desviaciones estándar por estrato, que son todas iguales a 10 en este caso, tenemos (los W_h son los del problema anterior):

$$n = \frac{\sum_{h=1}^{L} W_h \sigma_h^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h \sigma_h^2} = 75{,}6 \approx 76$$

Una vez hallado el tamaño de muestra, realizamos la afijación proporcional como sigue:

$$n_h = kN_h \text{ con } k = \frac{n}{N} = \frac{76}{310} = 0{,}245 \Rightarrow \begin{cases} n_1 = kN_1 = 0{,}245 \cdot 155 \cong 38 \\ n_2 = kN_2 = 0{,}245 \cdot 62 \cong 15 \\ n_3 = kN_3 = 0{,}245 \cdot 93 \cong 23 \end{cases}$$

4.7. Una empresa de publicidad quiere estimar la proporción de hogares en un municipio donde se ve cierto programa televisivo. El municipio tiene en total $N = 310$ hogares y es dividido en tres estratos (pueblo A, pueblo B y un área rural) de tamaños 155, 62 y 93 hogares, respectivamente. Datos de un estudio anterior indican que las proporciones de hogares donde se ve el programa pueden estimarse por 0,80, 0,25 y 0,30, respectivamente, en cada estrato. Además, el coste para obtener una observación es de 9 unidades monetarias para cualquiera de los pueblos y de 16 para el área rural. Hallar el tamaño de muestra n y su reparto entre los estratos para estimar la proporción poblacional de hogares donde se ve la televisión con un límite para el error de estimación igual a 0,1 y con un coste mínimo.

Resolver el problema suponiendo que las entrevistas se realizan por teléfono.

Resolver el problema suponiendo que las entrevistas se realizan por teléfono y la proporción de hogares donde se ve el programa televisivo es similar en cada uno de los tres estratos.

Observamos que, como el límite del error de estimación es 0,1, tenemos:

$$2\sqrt{\hat{P}_{st}} = 0{,}1 \Rightarrow \hat{P}_{st} = 0{,}0025$$

En la primera parte del problema se trata de buscar el tamaño de muestra necesario para estimar la proporción de hogares donde se ve el programa televisivo con un error de 0,1 y afijación óptima. Disponemos de los siguientes datos:

Estratos	*Tamaños*	c_i	$\hat{P}_h$
1	$N_1 = 155$	9	0.80
2	$N_2 = 62$	9	0.25
3	$N_3 = 93$	16	0,50

Como estamos en afijación óptima sin reposición, el tamaño de muestra necesario para cometer un error de muestreo unitario para estimar la proporción vendrá dado por:

$$V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right) - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}\right)\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

Aproximando las cuasivarianzas por $\hat{P}_h\hat{Q}_h = \hat{P}_h(1 - \hat{P}_h)$ por estrato tenemos:

$$n = \frac{\left(\sum_{h=1}^{L} \frac{N_h}{N}\sqrt{\hat{P}_h\hat{Q}_h} \Big/ \sqrt{c_h}\right)\left(\sum_{h=1}^{L} \frac{N_h}{N}\sqrt{\hat{P}_h\hat{Q}_h}\sqrt{c_h}\right)}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} \frac{N_h}{N}\hat{P}_h\hat{Q}_h} = 62{,}3 \approx 64$$

Ahora realizamos la afijación óptima como sigue:

$$n_h = n \cdot \frac{N_h\sqrt{P_hQ_h}/\sqrt{C_h}}{\sum_{h=1}^{L} N_h\sqrt{P_hQ_h}/\sqrt{C_h}} \Rightarrow \begin{cases} n_1 = 63 \cdot \dfrac{N_1\sqrt{P_1Q_1}/\sqrt{C_1}}{N_1\sqrt{P_1Q_1}/\sqrt{C_1} + N_2\sqrt{P_2Q_2}/\sqrt{C_2} + N_3\sqrt{P_3Q_3}/\sqrt{C_3}} \cong 31 \\ n_2 = 63 \cdot \dfrac{N_2\sigma_2}{N_1\sqrt{P_1Q_1}/\sqrt{C_1} + N_2\sqrt{P_2Q_2}/\sqrt{C_2} + N_3\sqrt{P_3Q_3}/\sqrt{C_3}} \cong 14 \\ n_3 = 63 \cdot \dfrac{N_3\sigma_3}{N_1\sqrt{P_1Q_1}/\sqrt{C_1} + N_2\sqrt{P_2Q_2}/\sqrt{C_2} + N_3\sqrt{P_3Q_3}/\sqrt{C_3}} \cong 18 \end{cases}$$

En caso de utilizar entrevista telefónica, los costes unitarios por estrato son iguales, con lo que la afijación óptima coincide con la afijación de mínima varianza. En este caso, el tamaño de muestra para cometer un error de muestreo unitario será:

$$e^2 = V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

Aproximando las cuasivarianzas por $\hat{P}_h\hat{Q}_h = \hat{P}_h(1 - \hat{P}_h)$ por estrato tenemos:

$$n = \frac{\left(\sum_{h=1}^{L} \frac{N_h}{N}\sqrt{\hat{P}_h\hat{Q}_h}\right)^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} \frac{N_h}{N}\hat{P}_h\hat{Q}_h} = 61{,}08 \approx 62$$

Una vez hallado el tamaño de muestra, realizamos la afijación de mínima varianza como sigue:

$$n_h = n \cdot \frac{N_h\sqrt{P_hQ_h}}{\sum_{h=1}^{L} N_h\sqrt{P_hQ_h}} \Rightarrow \begin{cases} n_1 = 62 \cdot \dfrac{N_1\sqrt{P_1Q_1}}{N_1\sqrt{P_1Q_1} + N_2\sqrt{P_2Q_2} + N_3\sqrt{P_3Q_3}} \cong 29 \\ n_2 = 62 \cdot \dfrac{N_2\sigma_2}{N_1\sqrt{P_1Q_1} + N_2\sqrt{P_2Q_2} + N_3\sqrt{P_3Q_3}} \cong 12 \\ n_3 = 62 \cdot \dfrac{N_3\sigma_3}{N_1\sqrt{P_1Q_1} + N_2\sqrt{P_2Q_2} + N_3\sqrt{P_3Q_3}} \cong 21 \end{cases}$$

Aunque el tamaño de muestra sigue siendo muy parecido al del apartado anterior, la afijación cambia, tomándose más observaciones del área rural ya que ahora no tienen un coste más alto.

Si, además de utilizar costes unitarios, suponemos que la variabilidad en los estratos es constante ($P_h \approx 0{,}4 \Rightarrow \sigma_h^2 = P_hQ_h = P_h(1-P_h) = 0{,}24$), podemos aproximar la afijación óptima y la de mínima varianza por la proporcional, ya que en este caso coinciden las tres. Entonces, el tamaño de muestra para cometer un error de muestreo unitario será:

$$e^2 = V(\bar{x}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\sum_{h=1}^{L} W_h S_h^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

Aproximando las cuasivarianzas por $\hat{P}_h\hat{Q}_h = \hat{P}_h(1 - \hat{P}_h)$ por estrato tenemos:

$$n = \frac{\sum_{h=1}^{L} \frac{N_h}{N}\hat{P}_h\hat{Q}_h}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} \frac{N_h}{N}\hat{P}_h\hat{Q}_h} = 73{,}3 \approx 74$$

Una vez hallado el tamaño de muestra, realizamos la afijación proporcional como sigue:

$$n_h = kN_h \text{ con } k = \frac{n}{N} = \frac{74}{310} = 0{,}238 \Rightarrow \begin{cases} n_1 = kN_1 = 0{,}238 \cdot 155 \cong 37 \\ n_2 = kN_2 = 0{,}238 \cdot 62 \cong 15 \\ n_3 = kN_3 = 0{,}238 \cdot 93 \cong 22 \end{cases}$$

4.8. Se trata de estimar el peso promedio de $N = 90$ conejos ($N_1 = 50$ machos y $N_2 = 40$ hembras) que han sido alimentados en cierta dieta. Los conejos se separan por sexo, por lo que el uso de muestreo aleatorio estratificado con dos estratos pareció apropiado. Para aproximar la variabilidad dentro de cada estrato, se pesó el conejo más pequeño y el más grande en cada estrato, y se halló que la amplitud de variación fue de 10 gramos para los machos y de 8 para las hembras. ¿Cuál es el tamaño de muestra necesario para estimar el peso promedio poblacional con un límite de 1 gramo para el error de estimación suponiendo que el costo de muestreo fue el mismo para ambos estratos?

Si suponemos los pesos con una distribución normal, la desviación estándar en cada estrato puede aproximarse por un cuarto de la amplitud de variación, es decir, $\sigma_1 = 10/4 = 2{,}5$ y $\sigma_2 = 8/4 = 2$.

Como los costes de muestreo son similares en los estratos, es lógico utilizar afijación de mínima varianza (que coincide con la óptima en este caso) y que siempre es más eficiente que la afijación proporcional. En este caso, el tamaño de muestra para cometer un error de muestreo unitario será:

$$e^2 = V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

Aproximando las cuasivarianzas por las desviaciones estándar por estrato tenemos:

$$n = \frac{\left(\sum_{h=1}^{L} \frac{N_h}{N}\sigma_h\right)^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} \frac{N_h}{N}\sigma_h^2} = 16{,}83 \approx 17$$

Una vez hallado el tamaño de muestra, realizamos la afijación de mínima varianza como sigue:

$$n_h = n \cdot \frac{N_h\sigma_h}{\sum_{h=1}^{L} N_h\sigma_h} \Rightarrow \begin{cases} n_1 = 17 \cdot \dfrac{N_1\sigma_1}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3} \cong 10 \\ n_2 = 17 \cdot \dfrac{N_2\sigma_2}{N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3} \cong 7 \end{cases}$$

4.9. Un mayorista del sector de la distribución de comestibles en una gran ciudad desea saber si la demanda es lo suficientemente grande para justificar la inclusión de un nuevo producto en sus existencias. Para tomar la decisión, planifica añadir este producto a una muestra de los almacenes a los que abastece para estimar el promedio de las ventas mensuales (variable X). El distribuidor suministra únicamente a cuatro grandes cadenas en la ciudad y, por conveniencia administrativa, decide utilizar muestreo aleatorio estratificado tomando cada cadena como un estrato. Hay 24 almacenes en el estrato 1, 36 en el estrato 2, 30 en el estrato 3 y 30 en el estrato 4 ($N_1 = 24$, $N_2 = 36$, $N_3 = 30$, $N_4 = 30$ y $N = 120$). El distribuidor tiene suficiente tiempo y dinero para obtener datos sobre ventas mensuales en una muestra de tamaño $n = 20$ almacenes. Dado que no tiene información previa respecto a las varianzas de los estratos y porque el coste del muestreo es el mismo en cada estrato, decide aplicar la afijación proporcional, con lo que el nuevo producto es introducido en cuatro almacenes elegidos al azar de la cadena 1, seis almacenes de la cadena 2, y 5 almacenes de cada una de las cadenas 3 y 4. Después de un mes, las ventas X presentan los resultados indicados en la tabla siguiente:

Estrato 1	Estrato 2	Estrato 3	Estrato 4
94	91	108	92
90	99	96	110
102	93	100	94
110	105	93	91
	111	93	113
	101		

Estimar las ventas promedio para el mes y fijar un límite para el error de estimación. Realizar la misma estimación y calcular el error suponiendo que se realiza muestreo aleatorio simple. Comentar los resultados.

Evidentemente, la afijación proporcional nos lleva a seleccionar cuatro almacenes elegidos al azar de la cadena 1, seis almacenes de la cadena 2, y 5 almacenes de cada una de las cadenas 3 y 4, ya que:

$$n_1 = n\left(\frac{N_1}{N}\right) = 20\left(\frac{24}{120}\right) = 4\,, \quad n_2 = 20\left(\frac{36}{120}\right) = 6\,, \quad n_3 = 20\left(\frac{30}{130}\right) = 5\,, \; n_4 = 20\left(\frac{30}{130}\right) = 5$$

De la tabla de ventas se deducen los siguientes valores:

Estratos →	*I*	*II*	*III*	*IV*
Medias ($\bar{x}_h$)	99	100	98	100
Cuasivarianzas ($\hat{S}_h^2$)	78,67	55,6	39,5	112,5

El estimador de la media será:

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h = \sum_{h=1}^{L} \frac{N_h}{N} \bar{x}_h = \frac{24}{120} 99 + \frac{36}{120} 100 + \frac{30}{120} 98 + \frac{30}{120} 100 = 99{,}3$$

Como la afijación es proporcional, tenemos:

$$\hat{V}(\bar{x}_{st}) = \frac{1}{n} \left(\sum_{h=1}^{L} W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2 = \frac{1}{n} \left(\sum_{h=1}^{L} \frac{N_h}{N} S_h \right)^2 - \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{N} S_h^2 = 2{,}93$$

Un intervalo de confianza al 95% para esta estimación será:

$$\bar{x}_{st} \pm 2\sqrt{\hat{V}(\bar{x}_{st})} = 99{,}3 \pm 2\sqrt{2{,}93} = 99{,}3 \pm 3{,}4$$

Si hubiésemos considerado muestreo aleatorio simple, el estimador de la media sería:

$$\bar{x} = \frac{1}{n} \sum_{h=1}^{L} X_i = \frac{1}{20} (94 + 90 + \cdots + 91 + 113) = 99{,}3$$

Su error de muestreo estimado será:

$$\hat{V}(\bar{x}) = (1 - f) \frac{\hat{S}^2}{n} = (1 - \frac{20}{120}) \frac{59{,}8}{20} = 2{,}49$$

Se observa que el error es menor en muestreo aleatorio simple con una ganancia en precisión dada por:

$$GP = \left(\frac{2{,}93}{2{,}49} - 1 \right) 100 = 17{,}67\%$$

La razón de que el muestreo estratificado proporcional haya sido peor que el aleatorio simple en un 17,67% radica en que las ventas varían fuertemente dentro de los almacenes de las distintas cadenas que conforman los estratos. Si observamos los valores de las cuasivarianzas en los distintos estratos vemos que varían mucho entre sí. Éste es un caso típico de mala aplicación de la afijación proporcional.

La posible solución a este problema podría haber sido la estratificación a partir de la cantidad de ventas, esto es, ubicando los almacenes con ventas mensuales bajas en un estrato, almacenes con ventas altas en otro, y así sucesivamente. De esta forma se conseguirían estratos muy homogéneos dentro de sí y heterogéneos entre sí, lo que disminuiría el error de estimación y aumentaría la ganancia en precisión del muestreo estratificado respecto del aleatorio simple.

4.10. La consejería de medio ambiente de una comunidad está realizando un estudio del número de personas X que utiliza las instalaciones de campings públicos. La comunidad tiene dos áreas para acampar, una localizada en las montañas y otra localizada a lo largo de la costa. La consejería desea estimar el número promedio de personas por camping y la proporción de campings que albergan personas de fuera de la comunidad durante un particular fin de semana, cuando se espera que todos los sitios estén ocupados. El número promedio de personas se va a estimar con un límite de 1 para el error de estimación, y la proporción de personas de fuera de la comunidad con un límite de 0,1. Las dos áreas para acampar forman convenientemente dos estratos, la localidad de la montaña como el estrato 1 y la localidad de la costa como el estrato 2. Se sabe que $N_1 = 120$ campings para acampar y $N_2 = 80$. Encuentre el tamaño de muestra y la asignación necesarios para lograr estos dos límites. Se supone que la consejería de medio ambiente conoce por experiencia que la mayoría de los campings contienen de 1 a 9 personas y que los costes de muestreo son los mismos en cada estrato.

Como los costes de muestreo son constantes en los estratos, utlizaremos afijación de mínima varianza (equivalente a la óptima en este caso). Además, como la desviación típica es alrededor de 1/4 de la amplitud de variación en una distribución normal, podemos suponer que su valor para el número de personas que ocupan los campings es constante en todos los campings y con valor $\sigma_i = (9 - 1)/4 = 2$.

En primer lugar observamos que, como el límite del error de estimación es 1 tenemos:

$$2\sqrt{\hat{V}(\bar{x}_{st})} = 1 \Rightarrow \hat{V}(\bar{x}_{st}) = 0{,}25$$

En afijación proporcional, el tamaño de muestra necesario para cometer un error de muestreo de 0,25 al estimar la media (promedio de personas por camping) podría estimarse como sigue:

$$e^2 = V(\bar{x}_{st}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h S_h^2 \Rightarrow n = \frac{\sum_{h=1}^{L} W_h S_h^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

Aproximando las cuasivarianzas por las desviaciones estándar por estrato, que son todas iguales a 2 en este caso, tenemos:

$$n = \frac{\sum_{h=1}^{L} \frac{N_h}{N}\sigma_h^2}{V(\bar{x}_{st}) + \frac{1}{N}\sum_{h=1}^{L} \frac{N_h}{N}\sigma_h^2} = 14{,}8 \approx 15$$

Una vez hallado el tamaño de muestra (15 campings), realizamos la afijación proporcional como sigue:

$$n_h = kN_h \text{ con } k = \frac{n}{N} = \frac{15}{200} = 0{,}075 \Rightarrow \begin{cases} n_1 = kN_1 = 0{,}075 \cdot 120 = 9 \\ n_2 = kN_2 = 0{,}075 \cdot 80 \cong 6 \end{cases}$$

Se estudiarán entonces 9 campings localizados en las montañas y 6 en la costa.

En el caso de la estimación de la proporción de ocupantes de fuera de la comunidad, no disponemos de estimaciones previas de las proporciones por estrato, lo que nos llevará a considerar $\hat{P}_1 = \hat{P}_2 = 0{,}5$ para obtener el tamaño de muestra máximo posible cuyo valor en afijación proporcional será:

$$n = \frac{\sum_{h=1}^{L} \frac{N_h}{N} \hat{P}_h \hat{Q}_h}{V(\bar{x}_{st}) + \frac{1}{N} \sum_{h=1}^{L} \frac{N_h}{N} \hat{P}_h \hat{Q}_h} = 67$$

Una vez hallado el tamaño de muestra, realizamos la afijación proporcional como sigue:

$$n_h = kN_h \text{ con } k = \frac{n}{N} = \frac{67}{200} = 0{,}335 \Rightarrow \begin{cases} n_1 = kN_1 = 0{,}335 \cdot 120 \cong 40 \\ n_2 = kN_2 = 0{,}238 \cdot 62 \cong 27 \end{cases}$$

Se estudiarán entonces 40 campings localizados en las montañas y 27 en la costa. Lógicamente se obtienen tamaños de muestra muy altos ya que nos hemos situado en el caso óptimo de precisión máxima.

4.11. Determinar el tamaño n de la muestra estratificada que con afijación de mínima varianza produzca la misma precisión que una muestra aleatoria simple (no estratificada) de tamaño n', para estimar la proporción P de una cierta clase en la población. Suponer en ambos casos muestreo con reposición y aplicar el resultado a los datos de la tabla con n'=1000.

	Estratos		
	I	*II*	*III*
W_h	0,2	0,3	0,5
P_h	0,5	0,6	0,4

Resolver el mismo problema para afijación proporcional y comparar resultados realizando los comentarios pertinentes.

Se trata de igualar la varianza del estimador de la proporción en muestreo estratificado con afijación de mínima varianza a la varianza del estimador de la proporción en el muestreo aleatorio simple en ambos casos con reposición. Se tiene:

$$V_{AS}(\hat{P}) = \frac{P(1-P)}{n'} \text{ y } V_{STMV}(\hat{P}) = \frac{\left(\sum_{h=1}^{3} W_h \sqrt{P_h(1-P_h)} \right)^2}{n}$$

Teniendo presente que $P = \sum W_h P_h$, se tiene el siguiente cuadro de datos:

Estratos	W_h	P_h	$1-P_h$	W_hP_h	$\sqrt{P_h(1-P_h)}$	$W_h\sqrt{P_h(1-P_h)}$
I	0,2	0,5	0,5	0,10	0,5	0,1
II	0,3	0,6	0,4	0,18	0,49	0,147
III	0,5	0,4	0,6	0,20	0,49	0,245
				$\sum_{h=1}^{3} W_hP_h = 48$		$\sum_{h=1}^{3} W_h\sqrt{P_h(1-P_h)} = 0{,}492$

Igualando las precisiones tenemos:

$$V_{AS}(\hat{P}) = V_{STMV}(\hat{P}) \Rightarrow \frac{P(1-P)}{n'} = \frac{\left(\sum_{h=1}^{3} W_h\sqrt{P_h(1-P_h)}\right)^2}{n} \Rightarrow$$

$$n = \frac{n'\left(\sum_{h=1}^{3} W_h\sqrt{P_h(1-P_h)}\right)^2}{P(1-P)} = \frac{1000\left(0{,}492\right)^2}{0{,}48(1-0{,}48)} = 970$$

Se obtiene un tamaño de muestra n = 970 en el muestreo estratificado con afijación de mínima varianza, que es ligeramente inferior al tamaño necesario en muestreo aleatorio simple n' = 1000. Existe entonces una ganancia en precisión por utilizar muestreo estratificado, pero es pequeña.

A continuación se iguala la varianza del estimador de la proporción en muestreo estratificado con afijación proporcional a la varianza del estimador de la proporción en el muestreo aleatorio simple, en ambos casos con reposición. Se tiene:

$$V_{AS}(\hat{P}) = \frac{P(1-P)}{n'} \text{ y } V_{STP}(\hat{P}) = \frac{\sum_{h=1}^{3} W_hP_h(1-P_h)}{n}$$

Igualando las precisiones tenemos:

$$V_{AS}(\hat{P}) = V_{STP}(\hat{P}) \Rightarrow \frac{P(1-P)}{n'} = \frac{\sum_{h=1}^{3} W_hP_h(1-P_h)}{n} \Rightarrow$$

$$n = \frac{n'\left(\sum_{h=1}^{3} W_hP_h(1-P_h)\right)}{P(1-P)} = \frac{1000\left(0{,}242\right)}{0{,}48(1-0{,}48)} = 970$$

Se obtiene un tamaño de muestra n = 970 en el muestreo estratificado con afijación proporcional, que es ligeramente inferior al tamaño necesario en muestreo aleatorio simple n'=1000. Existe entonces una ganancia en precisión por utilizar muestreo estratificado, pero es pequeña. Observamos que este tamaño de muestra con afijación proporcional coincide con el tamaño de muestra para afijación de mínima varianza, con lo que en este caso la precisión de ambos tipos de afijación es similar. Esto es debido a que las variabilidades por estrato $\sqrt{P_h(1-P_h)}$ son casi iguales (0,5, 0,49 y 0,49).

4.12. Se trata de estudiar el consumo anual de leche en una ciudad de 110000 habitantes. La población se divide en tres estratos por edades y se toman muestras aleatorias simples en cada uno de ellos. Se tienen los siguientes datos para el consumo anual de leche en litros:

Estratos	Tamaños poblacionales	Tamaños muestrales	Media muestral del consumo	Varianza muestral
Menores de 25 años	48000	1460	102,7	15876
Entre 25 y 50 años	38000	1160	71,4	48841
Más de 50 años	24000	1730	73,2	23409

Estimar la cantidad total de leche consumida al año entre los menores de 25 años, indicando el error de muestreo cometido, y calcular el tamaño muestral necesario para estimar el consumo medio de leche al año entre los mayores de 50 años, con un error de muestreo de 5 litros al 95% de confianza. Realizar una estimación por intervalos al 95% del consumo medio anual de leche por habitante.

Supongamos ahora que se multiplica por tres el tamaño de la muestra. Realizar la nueva afijación por los diferentes métodos para elegir el mejor.

Si se quiere estimar la proporción de personas entre 25 y 50 años que estarían dispuestas a comprar un producto lácteo de reciente aparición, ¿cuál sería el tamaño muestral necesario para estimar la proporción de personas entre 25 y 50 años que estarían dispuestas a comprar un producto lácteo de reciente aparición con un error de muestreo inferior al 5%? Comparar el resultado anterior con el obtenido cuando existe un estudio piloto que sugiere que dicha proporción será al menos del 60%.

Consideramos como variable X la cantidad de leche consumida anualmente por una persona. La primera pregunta del problema pide estimar un total poblacional dentro del primer estrato; por tanto, su estimador es el correspondiente a un muestreo aleatorio simple:

$$\hat{X}_1 = N_1 \bar{x}_1 = 48000(102{,}7) = 4929600 \text{ litros}$$

El error de muestreo estimado será:

$$\hat{V}(\hat{X}_1) = (1 - f_1)\frac{\hat{S}_1^2}{n} = (1 - \frac{1460}{48000})\frac{15876}{1460} \Rightarrow \hat{\sigma}(\hat{X}_1) = \sqrt{\hat{V}(\hat{X}_1)} = 155857{,}578 \text{ litros}$$

Par calcular el tamaño muestral necesario al estimar el consumo medio de leche en el tercer estrato con un error de muestreo fijado de 5 litros, será necesario un tamaño muestral igual a:

$$n_3 = \frac{\lambda_\alpha^2 N_3 S_3^2}{e_\alpha^2 N_3 + \lambda_\alpha^2 S_3^2} = \frac{1{,}96^2 (24000)(23409)}{5^2 (24000) + 1{,}96^2 (23409)} = 3128{,}25 \approx 3129$$

Para realizar la estimación por intervalos al 95% del consumo medio anual de leche por habitante en la ciudad, utilizaremos el estimador de la media global en el muestreo estratificado y, por tanto, el intervalo de confianza será:

$$\bar{x}_{st} \pm \lambda_\alpha \hat{\sigma}(\bar{x}_{st})$$

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{x}_h = \frac{1}{110000}\left(48000 * 102{,}7 + 38000 * 71{,}4 + 24000 * 73{,}2\right) = 85{,}451$$

$$\hat{\sigma}(\bar{x}_{st}) = \sqrt{\sum_{h=1}^{L} W_h^2 (1 - f_h)\frac{S_h^2}{n_h}} = \left(\frac{48000}{110000}\right)^2 * \left(1 - \frac{1460}{48000}\right) * \frac{15876}{1460} +$$

$$\left(\frac{38000}{110000}\right)^2 * \left(1 - \frac{1160}{38000}\right) * \frac{48841}{1160} + \left(\frac{24000}{110000}\right)^2 * \left(1 - \frac{1730}{24000}\right) * \frac{23409}{1730} = 2{,}73$$

Entonces:

$$\bar{x}_{st} \pm \lambda_\alpha \hat{\sigma}(\bar{x}_{st}) = 85{,}451 \pm 1{,}96 * 2{,}73 = \left[80{,}101;\quad 90{,}801\right]$$

Si triplicamos el tamaño de la muestra, el nuevo valor será 3(1460 + 1160 + 1730) = 13050 personas. A continuación realizamos las distintas afijaciones entre los estratos de este nuevo tamaño muestral.

Afijación uniforme

$W_h = 1/L = 1/3,\ h = 1, 2, 3\ \ n_1 = n_2 = n_3 = (1/L)n = 13050/3 = 4350$

Por tanto, de cada estrato se tomaría un muestra aleatoria simple de 4350 personas.

Afijación proporcional

$$W_h = \frac{N_h}{N}\ ,\ h = 1, 2, 3$$

$$n_1 = \frac{N_1}{N} n = \frac{48000}{110000} 13050 = 5694{,}5455 \approx 5695$$

$$n_2 = \frac{N_2}{N} n = \frac{38000}{110000} 13050 = 4508{,}1818 \approx 4508$$

$$n_3 = \frac{N_3}{N} n = \frac{24000}{110000} 13050 = 2847{,}2727 \approx 2847$$

Afijación de mínima varianza

$$n_h = \frac{N_h S_h}{\sum_{i=1}^{L} N_i S_i} n\ ,\ h = 1, 2, 3$$

$$\sum_{h=1}^{L} N_h S_h = 48000\sqrt{15876} + 38000\sqrt{48841} + 24000\sqrt{23409} = 18118000$$

$$n_1 = \frac{48000\sqrt{15876}}{18118000}13050 = 4356{,}2424 \approx 4356$$

$$n_2 = \frac{38000\sqrt{48841}}{18118000}13050 = 6048{,}8961 \approx 6049$$

$$n_3 = \frac{24000\sqrt{23409}}{18118000}13050 = 2644{,}8615 \approx 2645$$

Como la afijación de mínima varianza siempre supera a las demás, esta última es la afijación más eficiente entre los estratos.

Para resolver el último apartado utilizaremos muestreo aleatorio simple en el segundo estrato.

El tamaño muestral necesario para conseguir un error inferior a 0,05 al estimar la proporción con un coeficiente de confianza del 95% será una cantidad superior o igual a la siguiente:

$$n = \frac{\lambda_\alpha^2 N_2 p_2 q_2}{e_{p_3}^2 (N_2 - 1) + \lambda_\alpha^2 p_2 q_2} = \frac{1{,}96^2 (38000)(0{,}5)(0{,}5)}{0{,}05^2 (37999) + 1{,}96^2 (0{,}5)(0{,}5)} = 380{,}3251 \approx 381$$

Hemos supuesto que si no se tiene información sobre p_2 o q_2 tomamos $p_2 = q_2 = 0{,}5$, que es la situación de máxima variabilidad:

$$n = \frac{1{,}96^2 (38000)(0{,}5)(0{,}5)}{0{,}05^2 (37999) + 1{,}96^2 (0{,}5)(0{,}5)} = 380{,}3251 \approx 381$$

Si de la encuesta piloto se conoce que $p_2 \geq 0{,}6$, entonces tomaremos $p_2 = 0{,}6$ y $q_2 = 1 - 0{,}6 = 0{,}4$ con lo que:

$$n = \frac{1{,}96^2 (38000)(0{,}6)(0{,}4)}{0{,}05^2 (37999) + 1{,}96^2 (0{,}6)(0{,}4)} = 365{,}2583 \approx 366$$

Cuando no hay información sobe las proporciones poblaciones siempre nos situamos en la peor de las situaciones para nosotros en términos de coste, es decir, el caso en que más tamaño muestral se va a necesitar; sin embargo ésta es la situación de más precisión, es decir, que lo que se pierde en términos de coste se gana en términos de precisión.

Cualquier otro tamaño muestral obtenido para valores dados de la proporción poblacional distintos de 1/2 para cometer el mismo error de muestreo, será siempre menor.

14.13. Los 10000 trabajadores de una empresa fueron clasificados en tres grupos de edad, seleccionándose una muestra aleatoria simple en cada uno de ellos. Se obtuvieron las características siguientes para los tres grupos:

Gruposde edad	Número total de trabajadores	Número de trabajadores seleccionados	Salario mensual Media muestral	Salario mensual Desviación típica muestral	Número de contratos inferiores a 2 años
18-35	2900	666	120500	38000	375
36-50	4700	754	163000	35000	150
51-65	2400	580	195000	40000	90

Realizar una estimación por intervalos al 99% de confianza para el salario total percibido por los empleados más jóvenes. Hallar también la estimación del salario mensual medio de los diez mil trabajadores, así como su error de muestreo. Hallar el reparto muestral más eficiente en los distintos grupos de edad para estimar el salario mensual medio.

Realizar una estimación puntual de la proporción de trabajadores de la empresa cuyo contrato tiene una duración inferior a los dos años, indicando el error de muestreo cometido. Calcular el número de trabajadores que sería necesario seleccionar para que el error de muestreo no superase el 6% si se deseara estimar la proporción de trabajadores con más de 50 años que padecieron enfermedades por no cumplirse las normas de seguridad e higiene en sus puestos de trabajo.

Sea X el salario mensual de un trabajador. Para estimar el salario total repartido entre los trabajadores más jóvenes mediante un intervalo de confianza, utilizaremos la expresión correspondiente al muestreo aleatorio simple aplicada al primer estrato:

$$I_{\hat{X}_1} = \left[N_1\overline{X}_1 - \lambda_\alpha \sqrt{N_1^2(1-\frac{n_1}{N_1})\frac{S_1^2}{n_1}};\quad N_1\overline{X}_1 + \lambda_\alpha \sqrt{N_1^2(1-\frac{n_1}{N_1})\frac{S_1^2}{n_1}} \right]$$

El intervalo de confianza será entonces:

$$2900*120500 \pm 2{,}575\sqrt{2900^2\left(1-\frac{666}{2900}\right)\frac{38000^2}{66}} = \left[339799178{,}2;\quad 359100821{,}8\right]$$

Para estimar el salario medio de todos los trabajadores utilizamos el estimador del muestreo aleatorio estratificado:

$$\overline{x}_{st} = \sum_{h=1}^{L} W_h\overline{x}_h = 0{,}29*120500 + 0{,}47*163000 + 0{,}24*195000 = 158355$$

$$W_1 = \frac{N_1}{N} = \frac{2900}{10000} = 0{,}29\,,\quad W_2 = \frac{N_2}{N} = \frac{4700}{10000} = 0{,}47\,,\ W_3 = \frac{N_3}{N} = \frac{2400}{10000} = 0{,}24$$

El error de muestreo de la estimación anterior se calculará mediante:

$$\hat{\sigma}(\bar{x}_{st}) = \sqrt{\sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h}}$$

cuyo valor es:

$$\sqrt{0{,}29^2 \left(1 - \frac{666}{2900}\right) \frac{38000^2}{666} + 0{,}47^2 \left(1 - \frac{754}{4700}\right) \frac{35000^2}{754} + 0{,}24^2 \left(1 - \frac{580}{2400}\right) \frac{40000^2}{580}} = 749{,}85$$

La afijación más eficiente a realizar será la de mínima varianza, que siempre supera en precisión a las demás. Tenemos:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} n$$

$$\sum_{h=1}^{L} N_h S_h = 2900 * 38000 + 4700 * 35000 + 2400 * 40000 = 370700000$$

$$n_1 = \frac{2900 * 38000}{370700000} 2000 = 594{,}5508 \approx 595$$

$$n_2 = \frac{4700 * 35000}{370700000} 2000 = 887{,}5101 \approx 887$$

$$n_3 = \frac{2400 * 40000}{370700000} 2000 = 517{,}9390 \approx 518$$

Por tanto, el reparto muestral del enunciado no es el más eficiente.

Para estimar la proporción de trabajadores con contrato inferior a dos años, debemos obtener la estimación de la proporción poblacional en un muestreo aleatorio estratificado como sigue:

$$\hat{P}_{st} \sum_{h=1}^{L} W_h \hat{P}_h = 0{,}29 * 0{,}5631 + 0{,}47 * 0{,}1989 + 0{,}24 * 0{,}1552 = 0{,}2940$$

$$\hat{P}_1 = \frac{375}{666} = 0{,}5631\,,\ \hat{P}_2 = \frac{150}{754} = 0{,}1989\,,\ \hat{P}_3 = \frac{90}{580} = 0{,}1552$$

El error de muestreo de la estimación anterior será:

$$\hat{\sigma}(\hat{P}_{st}) = \sqrt{\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1} \frac{\hat{P}_h \hat{Q}_h}{n_h}}$$

cuyo valor es:

$$\sqrt{0{,}29^2\frac{2900-666}{2899}\frac{0{,}5631*0{,}4369}{666}+0{,}47^2\frac{4700-754}{4699}\frac{0{,}1989*0{,}8011}{754}+0{,}24^2\frac{2400-580}{2399}\frac{0{,}1552*0{,}8448}{580}}=0{,}008$$

En el último apartado hallamos el tamaño muestral necesario para estimar la proporción de trabajadores en el tercer estrato con un error de muestreo del 6%, que vendrá dado por:

$$n_3=\frac{\lambda_\alpha^2 N_3 p_3 q_3}{e_{p_3}^2\left(N_3-1\right)+\lambda_\alpha^2 p_3 q_3}=\frac{2{,}575^2*2400*0{,}5*0{,}5}{0{,}06^2\left(2399\right)+2{,}575^2*0{,}5*0{,}5}=386{,}4730\approx 387$$

Hemos supuesto que $p_3 = 0{,}5$ puesto que no se tiene información anterior sobre la proporción de trabajadores de más de 50 años que padecieron enfermedades por motivos laborales. Hemos llegado a que, para estimar esta proporción con un error de muestreo no superior al 6% habrá que seleccionar al menos 387 trabajadores entre el grupo de los mayores de 50 años.

4.14. Para estudiar el terreno agrícola de una comarca se consideraron tres zonas según su localización geográfica y en cada una de ellas, de forma independiente, se seleccionó, mediante un muestreo aleatorio simple, cierto número de fincas. Se tiene la siguiente información:

Zonas	Número total de fincas	Número de fincas seleccionadas	Superficie media muestral (Ha)	Desviación típica muestral (Ha)	Número de fincas barbecho
A	3200	380	28	3,5	124
B	5600	800	15	6,7	250
C	1200	200	45	8	17

Estimar puntualmente la superficie total del terreno agrícola en cada una de las zonas, así como su error de muestreo. Hallar los tamaños muestrales necesarios para realizar las estimaciones anteriores con unos errores de muestreo estimados inferiores a 1000 Ha y un coeficiente de confianza del 99%.

Realizar una estimación por intervalos al 99% de confianza de la superficie media de las fincas de la comarca y realizar la afijación más eficiente de la muestra anterior en las tres zonas para realizar la estimación de la superficie media.

Hallar también el tamaño muestral y la afijación que se debería haber realizado para estimar del modo más eficiente posible la superficie total del terreno agrícola de la comarca con un error de muestreo no superior a 1000 Ha y una confianza del 99%.

Estimar puntualmente el porcentaje global de fincas en barbecho y su error de muestreo.

Sea X la variable superficie de una finca de la comarca. Los estimadores dentro de cada zona podrán obtenerse a través de las fórmulas del muestreo aleatorio simple y los globales a partir de las del muestreo estratificado ya que las fincas de la comarca han sido divididas en tres zonas o estratos, y en cada una de ellas se ha realizado un muestreo aleatorio simple de forma independiente entre ellas.

Los estimadores puntuales de la superficie total del terreno agrícola en cada una de las zonas se calculan mediante $\hat{X}_h = N_h \bar{x}_h$ y su error de muestreo se calcula mediante:

$$\hat{\sigma}(\hat{X}_h) = \sqrt{N_h^2 \left(1 - \frac{n_h}{N_h}\right)} \frac{S_h}{\sqrt{n_h}} \quad , \; h = 1, 2, 3$$

Tenemos los siguientes resultados:

$$\hat{X}_1 = N_1 \bar{x}_1 = 3200 * 28 = 89600 \text{ Ha}$$

$$\hat{\sigma}(\hat{X}_1) = \sqrt{3200^2 \left(1 - \frac{380}{3200}\right)} \frac{3{,}5}{\sqrt{380}} = 539{,}35 \text{ Ha}$$

$$\hat{X}_2 = N_2 \bar{x}_2 = 5600 * 15 = 84000 \text{ Ha}$$

$$\hat{\sigma}(\hat{X}_1) = \sqrt{5600^2 \left(1 - \frac{800}{5600}\right)} \frac{6{,}7}{\sqrt{800}} = 1228{,}13 \text{ Ha}$$

$$\hat{X}_3 = N_3 \bar{x}_3 = 1200 * 45 = 54000 \text{ Ha}$$

$$\hat{\sigma}(\hat{X}_1) = \sqrt{1200^2 \left(1 - \frac{200}{1200}\right)} \frac{8}{\sqrt{200}} = 619{,}67 \text{ Ha}$$

Los tamaños muestrales necesarios para realizar las estimaciones anteriores con unos errores de muestreo estimados inferiores a 1000 Ha y una confianza del 99% se calculan en cada estrato mediante:

$$n_h = \frac{\lambda_\alpha^2 N_h^2 S_h^2}{e_{T_h}^2 + N_h \lambda_\alpha^2 S_h^2}, \; h = 1, 2, 3$$

Para los distintos estratos tendremos:

$$n_1 = \frac{5600^2 * 2{,}575^2 * 3{,}5^2}{1000^2 + 3200 * 2{,}575^2 * 3{,}5^2} = 660{,}1572 \approx 661$$

$$n_2 = \frac{5600^2 * 2{,}575^2 * 6{,}7^2}{1000^2 + 5600 * 2{,}575^2 * 6{,}7^2} = 3500{,}1310 \approx 3501$$

$$n_3 = \frac{1200^2 * 2{,}575^2 * 8^2}{1000^2 + 1200 * 2{,}575^2 * 8^2} = 404{,}8936 \approx 405$$

Para realizar la estimación por intervalos al 99% de la superficie media de las fincas de la comarca, utilizaremos el estimador de la media global en el muestreo estratificado y, por tanto, el intervalo de confianza será:

$$\bar{x}_{st} \pm \lambda_\alpha \hat{\sigma}(\bar{x}_{st})$$

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h = 0{,}32 * 28 + 0{,}56 * 15 + 0{,}12 * 45 = 22{,}76$$

$$N = \sum_{h=1}^{L} N_h = 3200 + 5600 + 1200 = 10000$$

$$W_1 = \frac{N_1}{N} = \frac{3200}{10000} = 0{,}32 \ ; \quad f_1 = \frac{n_1}{N_1} = \frac{380}{3200} = 0{,}1188$$

$$W_2 = \frac{N_2}{N} = \frac{5600}{10000} = 0{,}56 \ ; \quad f_2 = \frac{n_2}{N_2} = \frac{800}{5600} = 0{,}1429$$

$$W_3 = \frac{N_3}{N} = \frac{1200}{10000} = 0{,}12 \ ; \quad f_3 = \frac{n_3}{N_3} = \frac{200}{1200} = 0{,}1667$$

$$\hat{\sigma}(\bar{x}_{st}) = \sqrt{\sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h}} = \sqrt{0{,}32^2 (1 - 0{,}1188) \frac{3{,}5^2}{380} + 0{,}56^2 (1 - 0{,}1429) \frac{6{,}7^2}{800} + 0{,}12^2 (1 - 1667) \frac{8^2}{200}} = 0{,}147$$

Entonces:

$$\bar{x}_{st} \pm \lambda_\alpha \hat{\sigma}(\bar{x}_{st}) = 22{,}76 \pm 2{,}575 * 0{,}147 = [22{,}76 + 0{,}3805;\ 22{,}76 + 0{,}3805] = [22{,}3795;\ 23{,}1405]$$

La afijación más eficiente de la muestra anterior en las tres zonas para realizar la estimación de la superficie media será la afijación de mínima varianza definida por:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} n \quad , h = 1, \ldots, L$$

Tenemos:

$$\sum_{h=1}^{L} N_h S_h = 3200 * 3{,}5 + 5600 * 6{,}7 + 1200 * 8 = 58320$$

$$n = 380 + 800 + 200 = 1380$$

La afijación será la siguiente:

$$n_1 = \frac{3200 * 3,5}{58320} 1380 = 265,0205 \approx 265$$

$$n_2 = \frac{5600 * 6,7}{58320} 1380 = 887,8189 \approx 888$$

$$n_3 = \frac{1200 * 8}{58320} 1380 = 227,1605 \approx 227$$

En afijación de mínima varianza puede expresarse el tamaño muestral necesario para estimar el total con un error de muestreo dado e_α (1000 Ha) y un coeficiente de confianza adicional P_α, (99% $\Rightarrow \lambda_\alpha = 2,575$) mediante:

$$n = \frac{\sum_{h=1}^{L} \frac{N_h^2 S_h^2}{w_h}}{\frac{e_\alpha^2}{\lambda_\alpha^2} + \sum_{h=1}^{L} N_h S_h^2} \qquad w_h = \frac{N_h S_h}{\sum_{i=1}^{L} N_h S_h} \quad , h = 1, \ldots, L$$

$$w_1 = \frac{3200 * 3,5}{58320} = 0,1920$$

$$w_2 = \frac{5600 * 6,7}{58320} = 0,6433$$

$$w_3 = \frac{1200 * 8}{58320} = 0,1646$$

$$n = \frac{\frac{3200^2 * 3,5^2}{0,1920} + \frac{5600^2 * 6,7^2}{0,6433} + \frac{1200^2 * 8^2}{0,1646}}{\frac{1000^2}{2,575^2} + \left[3200 * 3,5^2 + 5600 * 6,7^2 + 1200 * 8^2\right]} = 6564,1970 \approx 6565$$

La afijación de los 6565 elementos muestrales en cada zona puede realizarse mediante:

$$n_h = w_h n \quad , h = 1, \ldots, L$$

$$n_1 = 0,1920 * 6565 = 1260,48 \approx 1261$$

$$n_2 = 0,6433 * 6565 = 4223,36 \approx 4223$$

$$n_3 = 0,1646 * 6565 = 1080,599 \approx 1081$$

Para estimar la proporción de fincas dedicadas al barbecho, definimos:

$$A_{hi} = \begin{cases} 1 & si\ la\ finca\ \ i\ del\ estrato\ h\ se\ destina\ a\ barbecho \\ 0 & en\ caso\ contrario \end{cases} \qquad h = 1, 2, 3$$

La proporción poblacional de fincas dedicadas al barbecho se estima mediante:

$$\hat{P}_{st} = \sum_{h=1}^{L} W_h \hat{P}_h = \sum_{h=1}^{L} \frac{N_h}{N} \hat{P}_h = \sum_{h=1}^{L} \frac{N_h}{N} \sum_{i=1}^{N_h} A_{ih} = 0{,}32\frac{124}{380} + 0{,}56\frac{250}{800} + 0{,}12\frac{17}{200} = 0{,}2896$$

$$\hat{P}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} A_{i1} = \frac{124}{380} = 0{,}3263$$

$$\hat{P}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} A_{i2} = \frac{250}{800} = 0{,}3125$$

$$\hat{P}_3 = \frac{1}{n_3} \sum_{i=1}^{n_3} A_{i3} = \frac{17}{200} = 0{,}085$$

El error de muestreo de este estimador se puede aproximar mediante:

$$\hat{\sigma}(\hat{P}_{st}) = \sqrt{\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1} \frac{\hat{P}_h \hat{Q}_h}{n_h}}$$

$$\hat{\sigma}(\hat{P}_{st}) = \sqrt{0{,}32^2 \frac{3200-380}{3200-1} \frac{0{,}3263 * 0{,}6737}{380} + 0{,}56^2 \frac{5600-800}{5600-1} \frac{0{,}3125 * 0{,}6875}{800} + 0{,}12^2 \frac{1200-200}{1200-1} \frac{0{,}085 * 0.915}{200}} = 0{,}0011$$

Un 28,96% de las fincas de la región están en barbecho, siendo el error de muestreo de esta estimación 0,0011.

4.15. En una ciudad turística de temporada con 10000 viviendas se desea conocer la proporción de viviendas en alquiler al menos una vez al año. Para realizar el estudio, se selecciona en cada uno de los tres barrios existentes una muestra aleatoria de viviendas de tamaño proporcional al número total de viviendas en cada uno. En el barrio A se seleccionaron 1050 viviendas, de las cuales había 800 en alquiler al menos un mes al año. En el barrio B se eligieron 900 viviendas, de las cuales había 600 en alquiler al menos un mes al año. En el barrio C se seleccionaron 1700 viviendas, de las cuales 1300 estaban en alquiler al menos un mes al año. Estimar la proporción de apartamentos que estarían dispuestos a ser alquilados al menos una vez al año y cuantificar el error de muestreo cometido.

Para estimar la proporción de viviendas en alquiler al menos una vez al año, definimos:

$$A_{hi} = \begin{cases} 1 & si\ la\ vivienda\ \ i\ del\ barrioo\ h\ se\ alquila\ al\ menos\ \ una\ vez\ al\ año \\ 0 & en\ \ caso\ \ contrario \end{cases} \qquad h = 1, 2, 3$$

La proporción de viviendas en alquiler al menos una vez al año se estima mediante:

$$\hat{P}_{st} = \sum_{h=1}^{L} W_h \hat{P}_h = \sum_{h=1}^{L} \frac{N_h}{N} \hat{P}_h \underbrace{=}_{\substack{\downarrow \\ \text{Afijación} \\ \text{proporcional}}} \sum_{h=1}^{L} \frac{n_h}{n} \hat{P}_h = \frac{1050}{3650}\frac{800}{1050} + \frac{900}{3650}\frac{600}{900} + \frac{1700}{3650}\frac{1300}{1700} = 0{,}7397$$

$$\hat{P}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} A_{i1} = \frac{800}{1050} = 0{,}7619,\ \hat{P}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} A_{i2} = \frac{600}{900} = 0{,}6667,\ \hat{P}_3 = \frac{1}{n_3}\sum_{i=1}^{n_3} A_{i3} = \frac{1300}{1700} = 0{,}7647$$

$$n = n_1 + n_2 + n_3 = 1050 + 900 + 1700 = 3650,\ N = 10000$$

Como la afijación es proporcional:

$$W_1 = \frac{N_1}{N} = \frac{n_1}{n} = \frac{1050}{3650} \Rightarrow N_1 = \frac{1050}{3650}10000 = 2877$$
$$W_2 = \frac{N_2}{N} = \frac{n_2}{n} = \frac{900}{3650} \Rightarrow N_2 = \frac{900}{3650}10000 = 2466$$
$$W_3 = \frac{N_3}{N} = \frac{n_3}{n} = \frac{1700}{3650} \Rightarrow N_3 = \frac{1700}{3650}10000 = 4658$$

El error de muestreo de este estimador se puede aproximar mediante:

$$\hat{\sigma}(\hat{P}_{st}) = \sqrt{\sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1}\frac{\hat{P}_h\hat{Q}_h}{n_h}}$$

$$\hat{\sigma}(\hat{P}_{st}) = \sqrt{0{,}32^2\frac{3200-380}{3200-1}\frac{0{,}3263*0{,}6737}{380} + 0{,}56^2\frac{5600-800}{5600-1}\frac{0{,}3125*0{,}6875}{800} + 0{,}12^2\frac{1200-200}{1200-1}\frac{0{,}085*0.915}{200}} = 0{,}0011$$

Un 28,96% de las fincas de la región está en barbecho, siendo el error de muestreo de esta estimación 0,0011.

4.16. Una gran empresa sabe que el 40% de las cuentas que recibe es al por mayor y el 60% es al por menor. Sin embargo, identificar las cuentas individuales sin consultar un archivo es complicado. Un auditor desea muestrear $n = 100$ de sus cuentas para estimar la cantidad promedio de las cuentas por cobrar de la empresa. Una muestra irrestricta aleatoria presenta 70% de cuentas al por mayor y un 30% de cuentas al por menor. Los datos son separados en cuentas al por mayor y cuentas al por menor después del muestreo, con los siguientes resultados en unidades monetarias:

Por mayor	Por menor
$n_1 = 70$	$n_2 = 30$
$\bar{y}_1 = 520$	$\bar{y}_2 = 280$
$\hat{S}_1 = 210$	$\hat{S}_2 = 90$

Estimar la cantidad promedio de las cuentas que recibe la empresa y fijar un límite para el error de estimación.

Como la proporción observada de cuentas al por mayor (0,7) está muy alejada de la proporción verdadera (0,4), la estratificación después de seleccionar la muestra irrestricta aleatoria (estratificación *a posteriori*) puede ser adecuada, lo cual puede también ser justificado porque n_1 y n_2 exceden de 20.

La cantidad promedio de cuentas que recibe la empresa se estima mediante:

$$\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h = \sum_{h=1}^{L} \frac{N_h}{N} \bar{x}_h = 0{,}4 * 520 + 0{,}6 * 280 = 376$$

El error de muestreo de la estimación anterior se calculará mediante:

$$\hat{\sigma}(\bar{x}_{st}) = \sqrt{\sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h}}$$

cuyo valor, omitiendo la corrección por población finita, es:

$$\sqrt{0{,}4^2 \frac{210^2}{70} + 0{,}6^2 \frac{90^2}{30}} = 14{,}07$$

El límite para el error de estimación al 95% será $2\hat{\sigma}(\bar{x}_{st}) \approx 28$, con lo que un intervalo de confianza al 95% para la estimación de la cantidad promedio de cuentas que recibe la empresa será 376 ± 28.

4.17. Un farmacéutico investiga el ingreso en caja obtenido por ventas a jubilados y al resto de sus clientes. Observa que el último mes ha vendido productos a 750 jubilados y 346 al resto de sus clientes. Como los jubilados suelen tener tratamientos particulares propios de enfermedades habituales en ellos, puede considerarse como un estrato homogéneo respecto de los productos que consumen. Lo mismo ocurre con el resto de los clientes. Como llevaría tiempo analizar cliente a cliente, se toma una muestra de 24 clientes y se estratifica *a posteriori* en función de si se trata de jubilados o no. El ingreso neto en euros por cada cliente de la muestra se presenta a continuación:

Cliente	*Ingreso*	*Cliente*	*Ingreso*	*Cliente*	*Ingreso*
Jubilado	271,3	*Normal*	173,69	*Jubilado*	277,67
Jubilado	301,29	*Normal*	133,24	*Normal*	171,89
Normal	163,17	*Jubilado*	275,8	*Normal*	165,22
Normal	141,72	*Normal*	246,48	*Jubilado*	235
Jubilado	367,94	*Normal*	176,7	*Normal*	181,2
Jubilado	328,63	*Jubilado*	292,09	*Normal*	177,37
Normal	179,7	*Normal*	187,52	*Normal*	161,37
Jubilado	337,77	*Jubilado*	349,79	*Normal*	215,76

Realizar una estimación del ingreso neto del farmacéutico y de su error de muestreo.

Como estamos ante un proceso de postestratificación, el número de jubilados y personas normales muestreadas son variables aleatorias con 24 valores. La cantidad ingresada por el farmacéutico se estima mediante:

$$\hat{X}'' = \sum_{h=1}^{L} N_h' \bar{x}_h = 750 * 303728 + 346 * 176{,}8 = 288968{,}8 \text{ euros}$$

La estimación de la varianza se calculará mediante:

$$\hat{V}(\hat{X}'') = \frac{N-n}{n} \sum_{h=1}^{L} N'_h \cdot \hat{S}_h^2 + \frac{N(N-n)}{n^2} \sum_{h=1}^{L} \hat{S}'^2_h (1 - f_h) = 71689746{,}68$$

El error relativo de muestreo será:

$$\hat{C}v(\hat{X}'') = \frac{\sqrt{71689746{,}68}}{288968{,}8} = 0{,}03 \rightarrow 3\%$$

EJERCICIOS PROPUESTOS

4.1. Sea X la variable salario anual en millones de unidades monetarias. Al medir la variable X sobre una población de 870 personas se obtiene la siguiente distribución de frecuencias:

Valores de X	2	3	4	7	10	12	16	20	25	30	35	50	60	100
Frecuencias (n_i)	20	30	60	100	150	200	120	80	50	20	18	10	8	4

Con el objeto de establecer pautas para futuras encuestas de salarios se estratifica la población utilizando dos métodos diferentes de estratificación. El método I consiste en realizar tres estratos según los criterios dados por $2 \le X \le 7$, $10 \le X \le 25$, $30 \le X \le 100$. El método II consiste en realizar tres estratos según los criterios dados por $2 \le X \le 10$, $12 \le X \le 35$, $50 \le X \le 100$. Se pide lo siguiente:

1º) Suponiendo muestreo con reposición y para un tamaño de muestra $n = 100$, realizar las afijaciones uniforme, proporcional y de mínima varianza para los dos métodos de estratificación. Comentar los resultados. Elegir el mejor método de estratificación y su tipo de afijación justificando la respuesta. Cuantificar la ganancia en precisión para el método y afijación elegidos respecto del muestreo aleatorio simple con reposición.

2º) Responder a las mismas cuestiones del apartado anterior suponiendo muestreo sin reposición. Comentar los resultados comparándolos con los del apartado anterior.

3º) Para la misma muestra de tamaño 100 realizar la afijación óptima para los dos métodos de estratificación, siendo los costes por unidad en cada estrato los siguientes: $C_{11} = 1$, $C_{21} = 16$, $C_{31} = 25$, $C_{12} = 4$, $C_{22} = 9$ y $C_{32} = 36$, donde C_{ij} = Coste por unidad en el estrato i según el método de estratificación j. Considerar muestreo sin reposición y con reposición y comparar los resultados. Para este tipo de afijación ¿cuál es el mejor método de estratificación? Razona la respuesta.

4º) En una encuesta de salarios posterior, ¿qué tamaño de muestra sería necesario para conseguir un error de muestreo de 0,5 al estimar la media salarial sin reposición y afijación de mínima varianza? ¿y si el muestreo es con reposición? Comentar los resultados.

5º) En una encuesta de salarios posterior ¿qué tamaño de muestra sería necesario para conseguir un error relativo de muestreo del 15% al 95% de coeficiente de confianza ($\lambda r\alpha = 1{,}96$) al estimar el total salarial con reposición y afijación proporcional. ¿Y si el muestreo es sin reposición? Comentar los resultados.

4.2. Se van a muestrear las familias de un pueblo para estimar la cantidad promedio de bienes por familia que se pueden convertir en dinero efectivo rápidamente. Las familias se estratifican en un estrato de renta alta y otro de renta baja. Se piensa que una casa en el estrato de renta alta tiene cerca de nueve veces más bienes que una casa en el estrato de renta baja, y se espera que S_h sea proporcional a la raíz cuadrada de la media del estrato. Se sabe que existen 4000 familias en el estrato de renta alta y 20000 familias en el estrato de renta baja. Se pide:

a) ¿Cómo se distribuiría de forma óptima entre los dos estratos una muestra de 1000 familias extraída de la población?

b) Si el objetivo es estimar la diferencia entre bienes por familia en ambos estratos ¿cómo debe distribuirse la muestra?

4.3. Consideramos un proceso de muestreo estratificado con afijación óptima en el que se define la función de coste total C de la siguiente forma:

$$C = c_0 + \sum_{h=1}^{L} c_h \sqrt{n_h}$$

donde c_0 representa un coste fijo dado y los c_h son también conocidos y representan el coste unitario en el estrato h ($h = 1, 2, ..., L$). Se pide:

1º Realizar la afijación de mínima varianza para un coste total C fijo al estimar la media poblacional y hallar la expresión general que nos da la varianza mínima.

2º Responder a las preguntas del apartado anterior considerando la extracción de una muestra estratificada de tamaño 1000 de una población de tamaño 10000 con los datos que se dan a continuación. Comparar los resultados con los que se obtendrían para afijación óptima con función de coste lineal y cuantificar la ganancia en precisión. Comentar los resultados.

Estrato	W_h	S_h	c_h
1	0,4	4	1
2	0,3	5	2
3	0,3	6	3

4.4. Supongamos conocidos los siguientes datos de una población dividida en tres estratos: $S_{12} = 9$, $S_{22} = 225$, $S_{32} = 1600$, $N_1 = 1000$, $N_2 = 600$, $N_3 = 200$, $C_1 = 1000$, $C_2 = 1200$ y $C_3 = 2000$. Se pide lo siguiente:

a) Determinar el coste de una muestra estratificada que proporciona un error relativo de muestreo de 5% para estimar la media considerando afijaciones proporcional, de mínima varianza y óptima, respectivamente. Se sabe que $\overline{X} = 22$ y que la función de coste es lineal. Comentar los resultados obtenidos para cada tipo de afijación y justificarlos.

b) Contestar a las mismas cuestiones del apartado anterior, pero con reposición, y comparar los resultados con los obtenidos en el apartado a). Justificar los resultados y comprobar que la afijación óptima y la de mínima varianza coinciden para costes unitarios.

CAPÍTULO

5

MUESTREO SISTEMÁTICO

OBJETIVOS

1. Presentar el concepto de muestreo sistemático.
2. Comprender las especificaciones del muestreo sistemático.
3. Analizar estimadores y errores en el muestreo sistemático.
4. Comprender el concepto de coeficiente de correlación intramuestral.
5. Analizar errores en función del coeficiente de correlación intramuestral.
6. Relacionar el muestreo sistemático con el muestreo aleatorio simple.
7. Relacionar el muestreo sistemático con el muestreo estratificado.
8. Comprender el concepto de coeficiente de correlación intraestratal.
9. Analizar errores en función del coeficiente de correlación intraestratal.
10. Realizar la estimación de varianzas.
11. Relacionar el muestreo sistemático con el muestreo por conglomerados.

ÍNDICE

1. Muestreo sistemático. Especificaciones.

2. Estimadores y varianzas.

3. Relación entre el muestreo sistemático y el muestreo aleatorio simple.

4. Relación entre el muestreo sistemático y el muestreo estratificado.

5. Estimación de varianzas.

6. Relación entre el muestreo sistemático y el muestreo por conglomerados.

7. Problemas resueltos.

8. Ejercicios propuestos

MUESTREO SISTEMÁTICO. ESPECIFICACIONES

Partimos de una población de tamaño N, y agrupamos sus elementos en n zonas (filas) de tamaño k ($N = nk$). Podríamos representar la población como sigue:

$i \setminus j$	1	2	3	$\cdots$	j	$\cdots$	k
1	u_{11}	u_{12}	u_{13}	$\cdots$	u_{1j}	$\cdots$	u_{1k}
2	u_{21}	u_{22}	u_{23}	$\cdots$	u_{1j}	$\cdots$	u_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	u_{i1}	u_{i2}	u_{i3}	$\cdots$	u_{1j}	$\cdots$	u_{ik}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	u_{n1}	u_{n2}	u_{n3}	$\cdots$	u_{nj}	$\cdots$	u_{nk}

A continuación se numeran los elementos de la tabla anterior de izquierda a derecha empezando por la primera unidad de la primera fila y pasando a la primera unidad de la fila siguiente cuando se agota cualquier fila. Tendríamos la siguiente estructura:

$i \setminus j$	1	2	3	$\cdots$	j	$\cdots$	k
1	u_1	u_2	u_3	$\cdots$	u_j	$\cdots$	u_k
2	u_{k+1}	u_{k+2}	u_{k+3}	$\cdots$	u_{k+j}	$\cdots$	u_{k+k}
3	u_{2k+1}	u_{2k+2}	u_{2k+3}	$\cdots$	u_{2k+j}		u_{2k+k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	$u_{(i-1)k+1}$	$u_{(i-1)k+2}$	$u_{(i-1)k+3}$	$\cdots$	$u_{(i-1)k+j}$	$\cdots$	$u_{(i-1)k+k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	$u_{(n-1)k+1}$	$u_{(n-1)k+2}$	$u_{(n-1)k+3}$	$\cdots$	$u_{(n-1)k+j}$	$\cdots$	$\underbrace{u_{(n-1)k+k}}_{u_N}$

Para extraer una muestra de tamaño n se elige al azar una unidad en la primera zona, y para seleccionar las $n - 1$ unidades restantes para la muestra se toma en cada zona la unidad que ocupa el mismo lugar dentro de su zona que el que ocupaba la primera unidad seleccionada dentro de la primera zona. Por ejemplo, si la unidad seleccionada para la muestra al azar en la primera zona es la tercera, se elegirán las $n - 1$ unidades restantes para la muestra tomando la tercera unidad de cada zona. Las muestras sistemáticas así obtenidas (columnas de la tabla anterior) suelen denominarse *muestras 1 en k*.

La probabilidad de seleccionar cualquier muestra será la probabilidad de elegir la unidad que la origina en la primera fila por muestreo aleatorio simple, es decir, $1/k$. Por tanto, el muestreo sistemático proporciona muestras equiprobables. Por otro lado, la probabilidad que tiene cualquier unidad de la población (de N unidades) de pertenecer a la muestra (de tamaño k) es $k/N = k/nk = 1/n$; por lo tanto, el muestreo sistemático es un tipo de muestreo con probabilidades iguales. Las muestras del espacio muestral pueden representarse como sigue:

$$\left(\tilde{u}_1\right) = \left\{u_1, u_{1+k}, \cdots u_{1+(n-1)k}\right\}$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\left(\tilde{u}_j\right) = \left\{u_j, u_{j+k}, \cdots u_{j+(n-1)k}\right\}$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\left(\tilde{u}_k\right) = \left\{u_k, u_{k+k}, \cdots u_{k+(n-1)k}\right\}$$

El muestreo sistemático extiende la muestra a toda la población, recoge el posible efecto de estratificación debido al orden en que figuran las unidades de la población (cada fila se puede considerar como un estrato), permite la consideración de conglomerados en la población (cada columna se puede considerar como un conglomerado), es fácil de aplicar y comprobar, no presenta problemas de cálculo algebraico y no precisa distinción entre reposición y no reposición. Además, si la disposición de los elementos en la población es aleatoria, la selección sistemática equivale a un muestreo aleatorio simple. Por último, el error de muestreo suele ser menor que en muestreo aleatorio simple o incluso que en estratificado.

Por otro lado, hay que tener presente la posibilidad de aumento de la varianza si existe periodicidad en la población y el problema teórico que se presenta en la estimación de varianzas debido a que no hay independencia en la selección de unidades en las distintas zonas, ya que la unidades extraídas en cada zona dependen de la seleccionada en la primera zona. En general sólo hay selección aleatoria para la primera unidad de la muestra.

ESTIMADORES Y VARIANZAS

Se utilizará el estimador lineal insesgado de Horwitz y Thompson porque el muestreo sistemático es sin reposición. En general sabemos que:

$$\hat{\theta}_{HT} = \sum_i^n \frac{Y_i}{\pi_i} \text{ estima insesgadamente el parámetro poblacional } \theta = \sum_i^n Y_i$$

En muestreo sistemático la probabilidad π_i de selección de un elemento poblacional cualquiera para la muestra será igual a la probabilidad de que resulte elegida la zona que lo contiene, esto es, $\pi_i = 1/k = n/(nk) = n/N$, luego podemos utilizar la forma general del estimador de Horwitz y Thompson para asegurar lo siguiente:

$$\hat{\theta} = \sum_i^n \sum_j^1 \frac{Y_{ij}}{\frac{1}{k}} \text{ estima insesgadamente el parámetro poblacional } \theta = \sum_i^n \sum_j^k Y_{ij}\,.$$

Si aplicamos la expresión anterior a las estimaciones del total, media, proporción y total de clase poblacionales, se obtienen los siguientes estimadores:

$$Total \rightarrow \theta = X \Rightarrow Y_{ij} = X_{ij} \Rightarrow \hat{X} = \sum_{i}^{n}\sum_{j=1}^{1}\frac{X_{ij}}{\frac{1}{k}} = \sum_{i=1}^{n}\underset{\downarrow}{k}\,X_{ij} = N.\frac{1}{n}\sum_{i=1}^{n}X_{ij} = N\bar{x}_j$$

$$Media \rightarrow \theta = \bar{X} \Rightarrow Y_{ij} = \frac{X_{ij}}{\underbrace{N}_{nk}} \Rightarrow \hat{\bar{X}} = \sum_{i}^{n}\sum_{j=1}^{1}\frac{\frac{X_{ij}}{nk}}{\frac{1}{k}} = \frac{1}{n}\sum_{i=1}^{n}X_{ij} = \bar{x}_j$$

$$Proporción \rightarrow \theta = P \Rightarrow Y_{ij} = \frac{A_{ij}}{nk} \Rightarrow \hat{P} = \sum_{i}^{n}\sum_{j=1}^{1}\frac{\frac{A_{ij}}{nk}}{\frac{1}{k}} = \frac{1}{n}\sum_{i=1}^{n}A_{ij} = \hat{P}_j$$

$$Totaldeclase \rightarrow \theta = X \Rightarrow Y_{ij} = A_{ij} \Rightarrow \hat{A} = \sum_{i}^{n}\sum_{j=1}^{1}\frac{A_{ij}}{\frac{1}{k}} = \sum_{i=1}^{n}\underset{\downarrow}{k}\,A_{ij} = N.\frac{1}{n}\sum_{i=1}^{n}A_{ij} = N\hat{P}_j$$

Hemos demostrado que un estimador lineal insesgado para la media poblacional es la media de la muestra sistemática obtenida, para la proporción poblacional es la proporción de la muestra sistemática, para el total poblacional es *N* veces el total de la muestra sistemática, y para el total de clase es *N* veces el total de clase muestral. Es decir, podemos escribir lo siguiente:

- $Total \rightarrow \hat{X} = N\bar{x}_j$
- $Media \rightarrow \hat{\bar{X}} = \bar{x}_j$
- $Proporción \rightarrow \hat{P} = \hat{P}_j$
- $Total\ de\ clase \rightarrow \hat{A} = N\hat{P}_j$

Varianzas de los estimadores

Definimos la cuasivarianza entre las *k* muestras posibles o cuasivarianza intermuestral como:

$$S_{bs}^2 = \frac{1}{k-1}\sum_{i}^{n}\sum_{j}^{k}\left(\bar{x}_j - \bar{X}\right)^2$$

y la cuasivarianza dentro de las muestras o cuasivarianza intramuestral como:

$$S_{ws}^2 = \frac{1}{N-k}\sum_{i}^{n}\sum_{j}^{k}\left(X_{ij} - \bar{x}_j\right)^2$$

Con lo que la descomposición de la suma de cuadrados para el análisis de la varianza poblacional permite escribir lo siguiente:

$$\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}\right)^2}_{(N-1)S^2} = \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \bar{x}_j\right)^2}_{(N-k)S_{ws}^2} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\bar{x}_j - \bar{X}\right)^2}_{(k-1)S_{bs}^2} \Rightarrow (N-1)S^2 = (N-k)S_{ws}^2 + (k-1)S_{bs}^2$$

A partir de la tabla del análisis de la varianza para la población que se presenta a continuación, pueden calcularse los errores de los estimadores.

Fuente de variación	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*
Entre muestas	$k-1$	$\sum_i^n \sum_j^k \left(\bar{x}_j - \bar{X}\right)^2$	S_{bs}^2
Dentro de muestras	$N-k$	$\sum_i^n \sum_j^k \left(X_{ij} - \bar{x}_j\right)^2$	S_{ws}^2
Total	$k-1+(N-k)=N-1$	$\sum_i^n \sum_j^k \left(X_{ij} - \bar{X}_j\right)$	S^2

$$V(\hat{\bar{X}}) = V\left(\bar{x}_j\right) = \left(1-f\right)\frac{S_{bs}^2}{n}, \quad V(\hat{X}) = V\left(N\bar{x}_j\right) = N^2 V(\bar{x}_j) = N^2\left(1-f\right)\frac{S_{bs}^2}{n}$$

$$V(\hat{P}) = V\left(\hat{P}_j\right) = \frac{1}{k}\sum_j^k \left(\hat{P}_j - P\right)^2 = \frac{1}{nk}\sum_i^n \sum_j^k \left(\hat{P}_j - P\right)^2 = \frac{1}{N}\sum_i^n \sum_j^k \left(\hat{P}_j - P\right)^2 = PQ - \frac{1}{k}\sum_j^k \hat{P}_j\hat{Q}_j$$

$$V(\hat{A}) = V\left(N\hat{P}_j\right) = N^2 V\left(\hat{P}_j\right) = N^2\frac{1}{k}\sum_j^k \left(\hat{P}_j - P\right)^2 = N\sum_i^n \sum_j^k \left(\hat{P}_j - P\right)^2 = N^2\left(PQ - \frac{1}{k}\sum_j^k \hat{P}_j\hat{Q}_j\right)$$

Un concepto interesante en muestreo sistemático es el ***coeficiente de correlación intramuestral*** ρ_w, que mide la interrelación entre las unidades dentro de las muestras. Lógicamente, esta interrelación debe ser lo más pequeña posible, ya que en el muestreo sistemático interesa la heterogeneidad intramuestral, con la finalidad de que una única muestra sistemática represente lo mejor posible a toda la población. Para que una muestra sistemática aspire a ser fiel espejo de toda la población ha de ser heterogénea, y la interrelación entre sus unidades ha de ser baja. Por lo tanto, inicialmente parece lógico que interesen valores muy pequeños del coeficiente de correlación intramuestral. La expresión matemática de ρ_w es la siguiente:

$$\rho_w = \frac{2\sum_j^k \sum_{i<z}^n \left(X_{ij} - \bar{X}\right)\left(X_{zj} - \bar{X}\right)}{N(n-1)\sigma^2}, \quad \sigma^2 = \frac{1}{nk}\sum_j^k \sum_i^n \left(X_{ij} - \bar{X}\right)^2 = \text{varianza poblacional}$$

La varianza de los estimadores puede expresarse en función de ρ_w. Para la media tenemos:

$$V\left(\bar{x}_j\right) = \frac{\sigma^2}{n}\left[1+(n-1)\rho_w\right] = \frac{N-1}{N}\frac{S^2}{n}\left[1+(n-1)\rho_w\right]$$

$$V(\hat{X}) = V\left(N\bar{x}_j\right) = N^2 V\left(\bar{x}_j\right) = N^2\frac{\sigma^2}{n}\left[1+(n-1)\rho_w\right] = N(N-1)\frac{S^2}{n}\left[1+(n-1)\rho_w\right]$$

$$V\left(\hat{P}_j\right) = \frac{PQ}{n}\left[1+(n-1)\rho_w\right] \qquad V\left(\hat{A}\right) = N^2\frac{PQ}{n}\left[1+(n-1)\rho_w\right]$$

Según esta expresión, la precisión del muestreo sistemático puede analizarse en función del coeficiente de correlación intramuestral, de tal modo que la precisión máxima se produce para $\rho_\omega = -1/(n\text{-}1)$, y la mínima para $\rho_\omega = 0$, *igualándose la precisión del muestreo sistemático con la del muestreo aleatorio simple para* $\rho_\omega = 0$. De esta forma, para valores de ρ_ω entre $-1/(n\text{-}1)$ y 0, el muestreo sistemático es más preciso que el aleatorio simple, y para valores de ρ_ω entre 0 y 1, el muestreo sistemático es menos preciso que el aleatorio simple. Por lo tanto, en cuanto a precisión, convienen valores negativos del coeficiente de correlación intraconglomerados ρ_ω.

RELACIÓN ENTRE MUESTREO SISTEMÁTICO Y MUESTREO ALEATORIO SIMPLE

El muestreo sistemático se ideó con la finalidad de mejorar el muestreo aleatorio simple. Pero habrá ocasiones en que esta mejora es máxima. Se demuestra que mientras más supera la cuasivarianza intramuestral S_{ws}^2 a la cuasivarianza poblacional S^2 el muestreo sistemático más gana en precisión al aleatorio simple.

El párrafo anterior puede interpretarse diciendo que el muestreo sistemático es más preciso que el aleatorio simple cuando la variabilidad dentro de muestras es superior a la variabilidad dentro de las unidades de la población. La precisión del muestreo sistemático coincide con la del aleatorio simple cuando $S_{ws}^2 = S^2$, es decir, cuando la variabilidad dentro de muestras es similar a la variabilidad dentro de las unidades de la población, y esto se da cuando la disposición de los elementos en la población es aleatoria.

RELACIÓN ENTRE MUESTREO SISTEMÁTICO Y MUESTREO ESTRATIFICADO

En el muestreo sistemático puede considerarse cada zona de k elementos consecutivos a partir del primero como un estrato; es decir, se puede dividir la población en n estratos constituidos cada uno de ellos por una fila de la tabla (k unidades) del cuadro en que hemos representado los elementos de la población numerados consecutivamente.

	1	j	k
1	x_1	x_j	x_k
2	x_{1+k}	x_{j+k}	x_{k+k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$x_{1+(i-1)k}\cdots$	$x_{j+(i-1)k}\cdots$	$x_{k+(i-1)k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_{1+(n-1)k}$	$x_{j+(n-1)k}$	$x_{k+(n-1)}k$

Obtener una muestra sistemática sería entonces equivalente a obtener una muestra estratificada con una unidad por estrato. Debe tenerse en cuenta, sin embargo, que en el muestreo estratificado aleatorio la selección se efectúa independientemente en cada estrato, mientras que en el muestreo sistemático todos los elementos seleccionados ocupan el mismo lugar o número de orden dentro de cada zona de k elementos, con la que no hay aleatoriedad de selección. Además, sería conveniente que las n zonas sistemáticas de k elementos cada una (estratos) sean lo más homogéneas posible dentro de ellas y heterogéneas entre ellas. Esta clasificación de los elementos de la población en n filas de k unidades cada una origina la siguiente tabla del análisis de la varianza poblacional:

Fuente de variación	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*
Entre estratos	$n-1$	$\sum_i^n \sum_j^k \left(\overline{X}_i - \overline{X}\right)^2$	S_{bst}^2
Dentro de estratos	$N-n$	$\sum_i^n \sum_j^k \left(X_{ij} - \overline{X}_i\right)^2$	S_{wst}^2
Total	$n-1+(N-n) = N-1$	$\sum_i^n \sum_j^k \left(X_{ij} - \overline{X}_j\right)$	S^2

Si definimos la cuasivarianza entre las n estratos posibles, o cuasivarianza interestratal como:

$$S_{bss}^2 = \frac{1}{n-1}\sum_{i}^{n}\sum_{j}^{k}\left(\overline{X}_i - \overline{X}\right)^2$$

y la cuasivarianza dentro de los estratos o cuasivarianza intraestratal como:

$$S_{wss}^2 = \frac{1}{N-n}\sum_{i}^{n}\sum_{j}^{k}\left(X_{ij} - \overline{X}_i^2\right)^2$$

tenemos:

$$\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \overline{X}\right)^2}_{(N-1)S^2} = \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \overline{X}_i\right)^2}_{(N-n)S_{wst}^2} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\overline{X}_i - \overline{X}\right)^2}_{(n-1)S_{bst}^2} \Rightarrow \ (N-1)S^2 = (N-n)S_{wst}^2 + (n-1)S_{bst}^2$$

Tenemos entonces que la varianza de la media puede expresarse como:

$$V\left(\hat{\overline{X}}\right) = V\left(\overline{x}_{st}\right) = \sum_{h}^{L} W_h^2 V\left(\overline{x}_h\right) = \sum_{i}^{n} W_i^2 V\left(\overline{x}_i\right) = \sum_{i}^{n}\frac{1}{n^2}\left(1-f_i\right)\cdot\frac{S_i^2}{n_i} = \frac{1}{n^2}\left(1-\frac{1}{k}\right)\sum_{i}^{n} S_i^2 =$$

$$\frac{1}{n^2}\left(1-\frac{1}{k}\right)\sum_{i}^{n}\frac{1}{k-1}\sum_{j}^{k}\left(X_{ij}-\overline{X}_i\right)^2 = \frac{1}{n^2 k}\underbrace{\sum_{i}^{n}\sum_{j}^{k}\left(X_{ij}-\overline{X}_i\right)^2}_{(N-n)S_{wst}^2} = \frac{N-n}{Nn}S_{wst}^2 = (1-f)\frac{S_{wst}^2}{n}$$

Si definimos ahora el ***coeficiente de correlación intraestratal*** $\rho_{\omega st}$ como el coeficiente de correlación lineal entre las desviaciones respecto de las medias de los estratos de todos los pares de valores que están en la misma muestra sistemática, su expresión puede calcularse de la siguiente forma:

$$\rho_{\omega st} = \frac{\operatorname{cov}\left(X_{ij};X_{zj}\right)}{\frac{1}{N}\sum_{j}^{k}\sum_{i=1}^{n}\left(X_{ij}-\overline{X}_i\right)^2} = \frac{\frac{1}{k\binom{n}{2}}\sum_{j}^{k}\sum_{i<z}^{n}\left(X_{ij}-\overline{X}_i\right)\left(X_{zj}-\overline{X}_z\right)}{\frac{1}{N}\underbrace{\sum_{j}^{k}\sum_{i=1}^{n}\left(X_{ij}-\overline{X}_i\right)^2}_{(N-n)S_{wst}^2}} = \frac{2\sum_{j}^{k}\sum_{i<z}^{n}\left(X_{ij}-\overline{X}_i\right)\left(X_{zj}-\overline{X}_z\right)}{n(n-1)(k-1)S_{wst}^2}$$

Se demuestra que la varianza del estimador de la media en función de $\rho_{\omega st}$ y $S_{\omega st}$ tiene la forma siguiente:

$$V(\hat{\overline{X}}) = V\left(\overline{x}_j\right) = (1-f)\frac{S_{wst}^2}{n}(1+(n-1)\rho_{\omega st})$$

y lo mismo se calcularían las varianzas del resto de los estimadores en función del coeficiente de correlación intraestratal $\rho_{\omega st}$.

La precisión máxima, que evidentemente se da cuando el error de muestreo es cero ($V(\bar{x}_j) = 0$), se produce si $(n-1)\rho_{\omega st} = -1$, luego se puede asegurar que la precisión máxima si:

$$V(\bar{x}_j) = 0 \Leftrightarrow \rho_{\omega st} = -\frac{1}{n-1}$$

La precisión mínima, que evidentemente se da cuando la varianza es máxima, se produce si $\rho_{\omega st} = 1$ (valor máximo de $\rho_{\omega st}$ que será el que efectivamente hace máxima $V(\bar{x}_j)$), luego se puede asegurar que PRECISIÓN MÍNIMA $\Leftrightarrow \rho_{\omega st} = 1$. Por otra parte:

$$\rho_{\omega st} = 0 \Rightarrow V(\bar{x}_j) = (1-f)\frac{S^2_{wst}}{n}$$

con lo que el muestreo sistemático coincide en precisión con el muestreo aleatorio estratificado considerando selección aleatoria independiente en cada estrato. De esta forma, $\rho_{\omega st}$ es en cierta forma una medida de la falta de aleatoriedad en la selección de unidades para la muestra en las distintas zonas sistemáticas (filas o estratos).

ESTIMACIÓN DE VARIANZAS

No podemos decir que en muestreo sistemático haya un método directo para la estimación de varianzas a partir de una muestra sistemática. Tenemos las siguientes situaciones:

a) ρ_ω próximo a cero o $S^2_{ws} = S^2$

Si el coeficiente de correlación intramuestral se aproxima a cero puede suponerse la población aleatoria y si $S^2_{ws} = S^2$ la precisión del aleatorio simple y el estratificado coinciden, con lo que la estimación de la varianza puede realizarse con la misma expresión que en muestreo aleatorio simple, es decir:

$$\hat{V}(\bar{x}) = \left(1-f\right)\cdot\frac{\hat{S}^2}{n}$$

siendo $\hat{S}^2$ la cuasivarianza de la muestra sistemática.

b) $\rho_{\omega st}$ próximo a cero

Si $\rho_{\omega st}$ se aproxima a cero se puede utilizar el muestreo sistemático como muestreo estratificado considerando cada zona sistemática como un estrato y seleccionando una muestra estratificada con una unidad por estrato. La razón de esta utilización es que la precisión del muestreo sistemático se iguala con la del muestreo aleatorio estratificado para $\rho_{\omega st} = 0$. En la práctica, lo que se hace es mezclar, antes de la selección, las $2k$ unidades de dos zonas en una única zona, con lo que se transforman las n zonas de k unidades cada una en $n/2$ zonas de $2k$ unidades cada una (si n es impar, para la zona que queda suelta se repite aleatoriamente un elemento de la muestra). Con este modelo se transforman las n zonas de k unidades en $n/2$ zonas de $2k$ unidades. Con ello se dispone de dos unidades muestrales por zona. Aplicando las fórmulas de muestreo estratificado tendremos:

$$\hat{V}\left(\bar{x}_{st}\right) = \sum_h^{\frac{n}{2}} W_h^2\left(1-f_h\right)\cdot\frac{\hat{S}_h^2}{n_h} = \sum_h^{\frac{n}{2}}\left(\frac{2}{n}\right)^2\left(1-f\right)\cdot\frac{\left(x_{h1}-x_{h2}\right)^2/2}{2} = \frac{1-f}{n^2}\sum_h^{\frac{n}{2}}\left(x_{h1}-x_{h2}\right)^2$$

c) Ni ρ_ω ni $\rho_{\omega st}$ están próximos a cero

En este caso utilizaremos alguno de los métodos especiales generales para la estimación de varianzas. Concretamente podemos utilizar el ***método de las muestras interpenetrantes***, que se utiliza cuando tenemos un conjunto de dos o más muestras, elegidas con el mismo esquema de muestreo (independientes o no) y tales que cada una proporcione una estimación válida del parámetro que se pretenda estimar con el mismo error de muestreo. Si las muestras son independientes es fácil obtener un estimador insesgado de la varianza del estimador. Para aplicar el método de las muestras interpenetrantes al muestreo sistemático supongamos que en vez de elegir una muestra sistemática de tamaño n para un solo valor j, $1 \le j \le k$, es decir, con un solo arranque aleatorio, obtenemos t muestras de tamaño n/t utilizando t arranques aleatorios. Estas muestras pueden considerarse independientes, ya que la elección del arranque es aleatoria en la primera zona sistemática.

Podemos formar un estimador combinado de la media poblacional basado en las medias de las t muestras (cada media muestral es un estimador insesgado de la misma media poblacional) definido como:

$$\bar{x}_c = \frac{1}{t}\sum_1^t \bar{x}_t$$

siendo el estimador insesgado de su varianza mediante la aplicación del método de las muestras interpenetrantes:

$$\hat{V}\left(\bar{x}_c\right) = \frac{1}{t(t-1)}\sum_i^t \bar{x}_i^2 - t\bar{x}_c^2 = \frac{1}{t(t-1)}\left(\sum_i^t \bar{x}_i^2 - \sum_i^t \bar{x}_c^2\right) = \frac{1}{t(t-1)}\sum_i^t\left(\bar{x}_i^2 - \bar{x}_c^2\right)$$

La fórmula puede multiplicarse también por (1-f). En particular para $t = 2$ tenemos:

$$\bar{x}_c = \frac{\bar{x}_1 + \bar{x}_2}{2} \Rightarrow \hat{V}\left(\bar{x}_c\right) = \bar{x}_1^2 - \left(\frac{\bar{x}_1 + \bar{x}_2}{2}\right)^2 + \bar{x}_2^2 - \left(\frac{\bar{x}_1 + \bar{x}_2}{2}\right)^2 = \frac{\left(\bar{x}_1 - \bar{x}_2\right)^2}{4}$$

Se observa que al aumentar el número de arranques aleatorios, manteniendo el mismo tamaño de muestra, la precisión obtenida se aproxima a la del muestreo aleatorio simple.

RELACIÓN ENTRE MUESTREO SISTEMÁTICO Y POR CONGLOMERADOS

En el muestreo sistemático puede considerarse cada columna de n elementos como un conglomerado; es decir, se puede dividir la población en k conglomerados constituidos cada uno de ellos por una columna de la tabla (n unidades). Obtener una muestra sistemática sería entonces equivalente a obtener una muestra por conglomerados de tamaño 1.

	1	j	k
1	x_1	x_j	x_k
2	x_{1+k}	x_{j+k}	x_{k+k}
⋮	⋮	⋮	⋮
i	$x_{1+(i-1)k}\cdots$	$x_{j+(i-1)k}\cdots$	$x_{k+(i-1)k}$
⋮	⋮	⋮	⋮
n	$x_{1+(n-1)k}$	$x_{j+(n-1)k}$	$x_{k+(n-1)}k$

PROBLEMAS RESUELTOS

5.1. En un proceso de fabricación de automóviles se trata de analizar la producción de piezas en serie de trece robots. Para ello se controlaron las piezas producidas por los trece robots en la primera hora de su funcionamiento y se obtuvo la siguiente distribución:

Nº de robot	1	2	3	4	5	6	7	8	9	10	11	12	13
Nº de piezas producidas	5	5	4	2	5	4	5	4	3	4	4	3	2

Con la finalidad de estimar el número de piezas defectuosas en el proceso de fabricación, se realiza un muestreo sistemático 1 en 5, es decir, se selecciona una de cada cinco piezas empezando por la primera pieza del primer robot hasta que se agoten sus piezas, para pasar a continuación a la primera pieza del segundo robot hasta que se agoten sus piezas, y así sucesivamente hasta que se agoten todas las piezas de todos los robots. Suponiendo que la primera pieza producida por cada robot es defectuosa y que todas las demás son correctas, se pide lo siguiente:

a) Calcular la varianza del estimador de la proporción de piezas defectuosas producidas por los robots y el valor del coeficiente de correlación intramuestral. ¿Existirá ganancia en precisión respecto de un muestreo irrestricto aleatorio con fracción de muestreo del 20%? ¿Por qué? Cuantificarla. Realizar la tabla del análisis de la varianza para la producción total.

b) Estimar la varianza para cada muestra sistemática posible según nuestro procedimiento de muestreo. ¿Con qué muestra sistemática nos quedaremos que represente mejor a toda la producción? ¿Existirá ganancia en precisión si se estiman las varianzas utilizando estratificación? Dar la estimación de la proporción de piezas defectuosas producidas por los robots.

Si definimos una variable dicotómica *A* a la que asignamos el valor 1 para las piezas defectuosas y el valor 0 para las piezas correctas, y clasificamos las 50 piezas en 10 filas de 5 piezas cada una (muestreo sistemático 1 en 5) siguiendo el orden del enunciado del problema, tendremos la tabla de la Figura 5-1.

A continuación, se construye la tabla del análisis de la varianza para la población (producción total) utilizando Excel. Como estamos clasificando los datos en 5 grupos (columnas), utilizaremos una variable *G*, que clasificará los valores de *A* (ceros o unos) por grupos (por columnas). Introducimos los valores de *G* en columnas de la hoja de cálculo de Excel y elegimos *Análisis de la varianza de un factor* en la opción *Análisis de datos* del menú *Herramientas*, rellenando su pantalla de entrada como se indica en la Figura 5-2. La Figura 5-3 presenta los resultados.

1	0	0	0	0	1/5
1	0	0	0	0	1/5
1	0	0	0	1	2/5
0	1	0	0	0	1/5
0	1	0	0	0	1/5
1	0	0	0	0	1/5
1	0	0	0	1	2/5
0	0	1	0	0	1/5
0	1	0	0	0	1/5
1	0	0	1	0	2/5
6/10	3/10	1/10	1/10	2/10	13/50

Figura 5-1

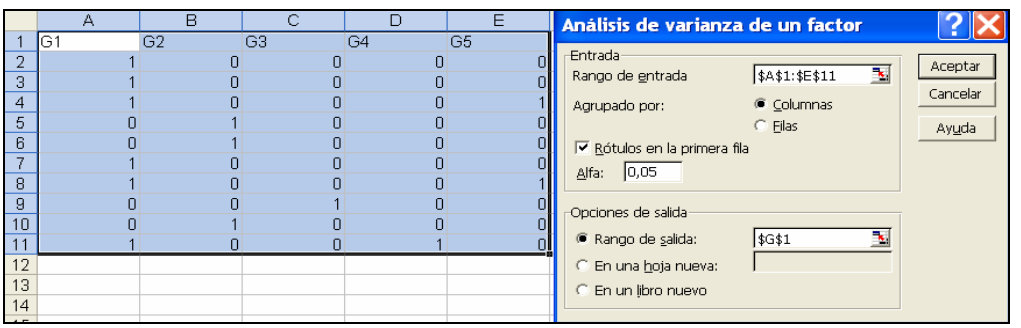


Figura 5-2

	G	H	I	J	K	L	M
1	Análisis de varianza de un factor						
2							
3	RESUMEN						
4	*Grupos*	*Cuenta*	*Suma*	*Promedio*	*Varianza*		
5	G1	10	6	0,6	0,266667		
6	G2	10	3	0,3	0,233333		
7	G3	10	1	0,1	0,1		
8	G4	10	1	0,1	0,1		
9	G5	10	2	0,2	0,177778		
10							
11							
12	ANÁLISIS DE VARIANZA						
13	*Origen de las variaciones*	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
14	Entre grupos	1,72	4	0,43	2,449367	0,05972034	2,578737224
15	Dentro de los grupos	7,9	45	0,175555556			
16							
17	Total	9,62	49				
18							

Figura 5-3

Con la notación que utilizamos habitualmente, la tabla ANOVA será:

Fuente	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*
Entre	$k-1=5-1=4$	$\sum_{i}^{n}\sum_{j}^{k}\left(\bar{x}_j-\bar{X}\right)^2=1{,}72$	$S_{bs}^2=1{,}72/4=0{,}43$
Dentro	$N-k=50-5=45$	$\sum_{i}^{n}\sum_{j}^{k}\left(X_{ij}-\bar{x}_j\right)^2=7{,}9$	$S_{ws}^2=7{,}9/45=0{,}1755$
Total	$N-1=50-1=49$	$\sum_{i}^{n}\sum_{j}^{k}\left(X_{ij}-\bar{X}_j\right)=9{,}62$	$S^2=9{,}62/49=0{,}1963$

Conocida esta tabla, pueden realizarse ya todos los cálculos. La varianza del estimador de la proporción puede calcularse como:

$$V(\hat{P}) = V\left(\hat{P}_j\right) = \left(1 - \frac{n}{N}\right)\frac{S_{bs}^2}{n} = (1-f)\frac{S_{bs}^2}{n} = (1 - \frac{1}{5})\frac{0{,}43}{10} = 0{,}0344$$

La varianza para el estimador del total de clase será:

$$V(\hat{A}) = V\left(N\hat{P}_j\right) = N^2 V(\hat{P}_j) = N^2(1-f)\frac{S_{bs}^2}{n} = 50^2 * 0{,}0344 = 86$$

Del valor de la varianza puede deducirse el valor del coeficiente de correlación intramuestral a través de la fórmula $V(\bar{x}_j) = \frac{\sigma^2}{n}(1 + (n-1)\rho_\omega)$. Tendremos:

$$0{,}0344 = \frac{\frac{49}{50}0{,}1963}{10}(1 + (10-1)\rho_\omega) \Rightarrow \rho_\omega = 0{,}0875$$

Se observa un valor de ρ_ω muy cercano a cero, lo que indica que el muestreo sistemático va a tener una precisión muy cercana a la del aleatorio simple en la estimación de la proporción de piezas defectuosas. Esto concuerda con el hecho de que S^2 y S^2_{ws} también tienen valores muy cercanos. Concretamente $S^2 = 0{,}1963 > S^2_{ws} = 0{,}1755$, lo que indica que es más preciso el muestreo aleatorio simple. La varianza del estimador de la proporción en el muestreo aleatorio simple es (1–1/5)0,1963/10 = 0,0157, lo que indica que la ganancia en precisión del aleatorio simple será (0,0344 – 0,0157)/0,0344 = 54,3%.

Dado el valor del coeficiente de correlación intramuestral, muy cercano a cero, podemos estimar varianzas mediante la fórmula del muestreo aleatorio simple. Se tiene:

$$\hat{V}\left(\hat{P}_1\right) = (1-f)\frac{\hat{S}_1^{\,2}}{n} = (1-f)\frac{\hat{P}_1\hat{Q}_1}{n-1} = \left(1-\frac{1}{5}\right)\frac{\frac{6}{10_1}\left(1-\frac{6}{10}\right)}{10-1} = 0{,}0213$$

$$\hat{V}\left(\hat{P}_2\right) = (1-f)\frac{\hat{S}_2^{\,2}}{n} = (1-f)\frac{\hat{P}_2\hat{Q}_2}{n-1} = \left(1-\frac{1}{5}\right)\frac{\frac{3}{10_1}\left(1-\frac{3}{10}\right)}{10-1} = 0{,}0186$$

$$\hat{V}\left(\hat{P}_3\right) = (1-f)\frac{\hat{S}_3^{\,2}}{n} = (1-f)\frac{\hat{P}_3\hat{Q}_3}{n-1} = \left(1-\frac{1}{5}\right)\frac{\frac{1}{10_1}\left(1-\frac{1}{10}\right)}{10-1} = 0{,}008$$

$$\hat{V}\left(\hat{P}_4\right) = \hat{V}\left(\hat{P}_3\right) = 0{,}008$$

$$\hat{V}\left(\hat{P}_5\right) = (1-f)\frac{\hat{S}_5^{\,2}}{n} = (1-f)\frac{\hat{P}_5\hat{Q}_5}{n-1} = \left(1-\frac{1}{5}\right)\frac{\frac{2}{10_1}\left(1-\frac{2}{10}\right)}{10-1} = 0{,}0142$$

Según estos resultados la muestras más precisas son la tercera y la cuarta.

También podemos estimar la varianza a partir del muestreo estratificado, agrupando las 10 filas (estratos) de la población en grupos de 2, y considerando cada dos filas como un estrato del que seleccionamos dos unidades para la muestra. Tendremos:

$$\hat{V}\left(\hat{P}_1\right) = \frac{1-f}{n^2}\sum_{h}^{\frac{n}{2}}\left(x_{h1}-x_{h2}\right)^2 = \frac{1-0{,}2}{10^2}\left[(1-1)^2+(1-0)^2+(0-1)^2+(1-0)^2+(0-1)^2\right] = 0{,}032$$

$$\hat{V}\left(\hat{P}_2\right) = \frac{1-f}{n^2}\sum_{h}^{\frac{n}{2}}\left(x_{h1}-x_{h2}\right)^2 = \frac{1-0{,}2}{10^2}\left[(0-0)^2+(0-1)^2+(1-0)^2+(0-0)^2+(1-0)^2\right] = 0{,}024$$

$$\hat{V}\left(\hat{P}_3\right) = 0{,}008(0-1)^2 = 0{,}008 = \hat{V}\left(\hat{P}_4\right) \qquad \hat{V}\left(\hat{P}_5\right) = 0{,}008\left[(1-0)^2+(1-0)^2\right] = 0{,}016$$

Las mejores muestras según el método del muestreo estratificado también resultan ser la tercera y la cuarta, y además coinciden en varianza con el método anterior. Para las restantes muestras se observa ganancia en precisión del método de estimación utilizando la fórmula del muestreo aleatorio simple. La proporción estimada de piezas defectuosas producidas será la derivada de la 3ª o 4ª muestra, esto es: $\hat{P} = \hat{P}_3 = \hat{P}_4 = 1/10$; es decir que se estima un 10% de producción defectuosa.

5.2. En una población de 8 tipos de maletines de herramientas medimos el número de elementos importantes que faltan para considerarse de primera calidad:

m_i	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8
X_i	1	3	5	2	4	6	2	7

Se realiza muestreo sistemático 1 en 2 y se pide:

a) Calcular las varianzas de los estimadores insesgados del total y de la media de elementos importantes ausentes en los maletines. Utilizar adicionalmente la relación entre muestreo sistemático y estratificado.

b) Estimar dichas varianzas y comparar la precisión de este tipo de muestreo con la del muestreo aleatorio simple. Seleccionar la muestra más precisa.

Como se trata de un muestreo sistemático 1 en 2 y $N = 8$, habrá dos muestras sistemáticas posibles de tamaño 4 (columnas). Dividiremos entonces la población en 4 zonas (filas) de 2 elementos cada una de la forma:

1	3	2
5	2	3,5
4	6	5
2	7	4,5
3	4,5	3,75

$$\sum_{i=1}^{3}\sum_{j=1}^{3}(\bar{x}_j-\bar{X})^2 = 4\left[(3-3{,}75)^2+(4{,}5-3{,}75)^2\right] = 4{,}5$$

$$\sum_{i=1}^{3}\sum_{j=1}^{3}(X_{ij}-\bar{x}_j)^2 = (1-3)^2+(5-3)^2+\ldots+(6-4{,}5)^2+(7-4{,}5)^2 = 27$$

$$\sum_{i=1}^{3}\sum_{j=1}^{3}(X_{ij}-\bar{X})^2 = (1-3{,}75)^2+(5-3{,}75)^2+\ldots+(7-3{,}75)^2 = 31{,}5$$

Hemos creado un cuadro con las muestras sistemáticas como columnas, colocando una fila adicional inferior con las medias de las columnas y una columna adicional a la derecha con las medias de las filas.

A continuación, se construye la tabla del análisis de la varianza para la población utilizando Excel. Como estamos clasificando los datos en 2 grupos (columnas), utilizaremos las variables M_1 y M_2, que recogen los valores de las dos columnas. A continuación elegimos *Análisis de la varianza de un factor* en la opción *Análisis de datos* del menú *Herramientas*, rellenando su pantalla de entrada como se indica en la Figura 5-4. La Figura 5-5 presenta los resultados.

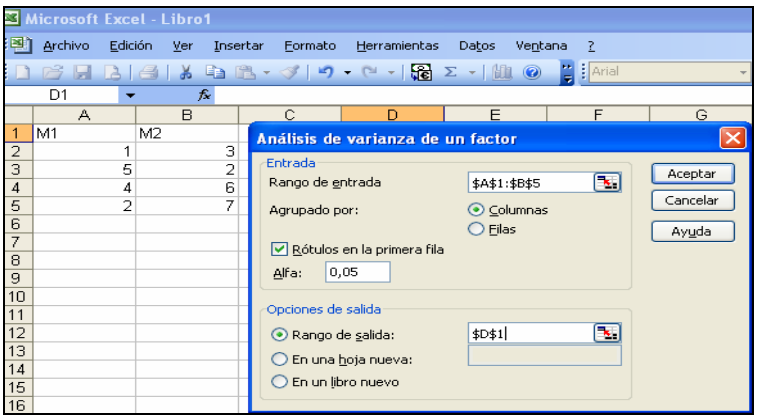


Figura 5-4

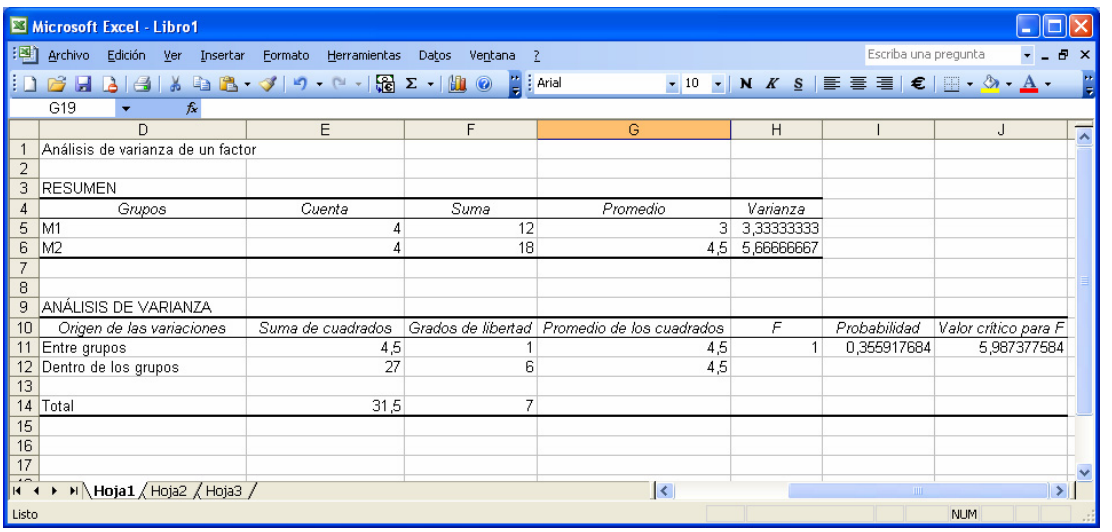


Figura 5-5

Mediante Excel se ha hallado la siguiente tabla del análisis de la varianza:

Fuente de variación	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*
Entre muestras	$k-1=2-1=1$	$\sum_i^n \sum_j^k \left(\bar{x}_j - \bar{X}\right)^2 = 4{,}5$	$S_{bs}^2 = 4{,}5/1 = 4{,}5$
Dentro de muestras	$N-k=8-2=6$	$\sum_i^n \sum_j^k \left(X_{ij} - \bar{x}_j\right)^2 = 27$	$S_{ws}^2 = 27/6 = 4{,}5$
Total	$N-1=8-1=7$	$\sum_i^n \sum_j^k \left(X_{ij} - \bar{X}_j\right) = 31{,}5$	$S^2 = 31{,}5/7 = 4{,}5$

Conocida esta tabla pueden realizarse ya todos los cálculos.

$$V(\hat{\bar{X}}) = V(\bar{x}_j) = \frac{1}{k}\sum_j^k (\bar{x}_j - \bar{X})^2 = \frac{1}{2}\left[(3-3{,}75)^2 + (3{,}5-3{,}75)^2\right] = 0{,}5625$$

La varianza del estimador de la media también puede calcularse como:

$$V(\hat{\bar{X}}) = V(\bar{x}_j) = \left(1-\frac{n}{N}\right)\frac{S_{bs}^2}{n} = (1-f)\frac{S_{bs}^2}{n} = (1-\frac{1}{2})\frac{4{,}5}{4} = 0{,}5625$$

La varianza para el estimador del total será:

$$V(\hat{X}) = V(N\bar{x}_j) = N^2 V(\bar{x}_j) = N^2(1-f)\frac{S_{bs}^2}{n} = 8^2 \cdot 0{,}5625 = 36$$

El cálculo de la varianza también puede realizarse a través del valor del coeficiente de correlación intramuestral como $V(\bar{x}_j) = \frac{\sigma^2}{n}(1+(n-1)\rho_\omega)$. Tenemos:

$$\rho_w = \frac{2\sum_j^k\sum_{i<z}^n (X_{ij}-\bar{X})(X_{zj}-\bar{X})}{N(n-1)\sigma^2} = \frac{2\sum_j^k\sum_{i<z}^n (X_{ij}-\bar{X})(X_{zj}-\bar{X})}{(N-1)(n-1)S^2} = -0{,}14285$$

Tendremos entonces:

$$V(\bar{x}_j) = \frac{\sigma^2}{n}(1+(n-1)\rho_\omega) = \frac{\frac{7}{8}4{,}5}{4}(1+3(-0{,}14285)) = 0{,}5625$$

Ahora surge el problema de estimar las varianzas. Para ello observamos en primer lugar que $S^2_{ws} = 4{,}5 = S^2$, por lo que la precisión en muestreo aleatorio simple coincide con la precisión del muestreo sistemático, y podremos utilizar la fórmula del muestreo aleatorio simple para estimar varianzas. Por otra parte, el valor del coeficiente de correlación intramuestral ρ_ω indica que la precisión del muestreo sistemático es buena, ya que éste es muy bajo y además es negativo. Al ser negativo vemos que no existe interrelación dentro de las muestras, esto es, que las muestras tienden a ser heterogéneas dentro de sí, lo cual es muy conveniente en muestreo sistemático a la vista de que la muestra ha de representar fielmente a toda una población que se supone heterogénea.

Para estimar la varianza de la media podemos utilizar la fórmula del muestreo aleatorio simple, ya que en este problema coincide en precisión con el sistemático. Tendremos los siguientes resultados para cada una de las dos muestras:

$$\hat{V}(\bar{x}_1) = (1-f)\cdot\frac{\hat{S}_1^{\,2}}{n} = \left(1-\frac{1}{2}\right)\left(\frac{1}{3}\left[(1-3)^2 + (5-3)^2 + (4-3)^2 + (2-3)^2\right]\Big/4\right) = 0{,}41$$

$$\hat{V}(\bar{x}_2) = (1-f)\cdot\frac{\hat{S}_2^{\,2}}{n} = \left(1-\frac{1}{2}\right)\left(\frac{1}{3}\left[(3-4{,}5)^2 + (2-4{,}5)^2 + (6-4{,}5)^2 + (7-4{,}5)^2\right]\Big/4\right) = 0{,}71$$

La mejor muestra sistemática resulta ser la primera, pues es la que presenta menor varianza.

También podemos tratar este problema desde el enfoque de la equivalencia entre muestreo estratificado y muestreo sistemático.

Consideramos ahora cada una de las 4 zonas (filas) como un estrato de 2 unidades. Tenemos entonces dividida la población en 4 estratos de 2 unidades cada uno, de modo que la muestra sistemática consta de una unidad por estrato, que de forma general no es elegida aleatoriamente dentro del mismo. Esta clasificación de los elementos de la población en 4 filas de 2 unidades cada una origina una tabla del análisis de la varianza para la población que puede calcularse a través de Excel. Como estamos clasificando los datos en 4 filas (estratos), utilizaremos las variables M_2 a M_5, que recogen los valores de las cuatro filas. A continuación elegimos *Análisis de la varianza de un factor* en la opción *Análisis de datos* del menú *Herramientas*, rellenando su pantalla de entrada como se indica en la Figura 5-6. La Figura 5-7 presenta los resultados.

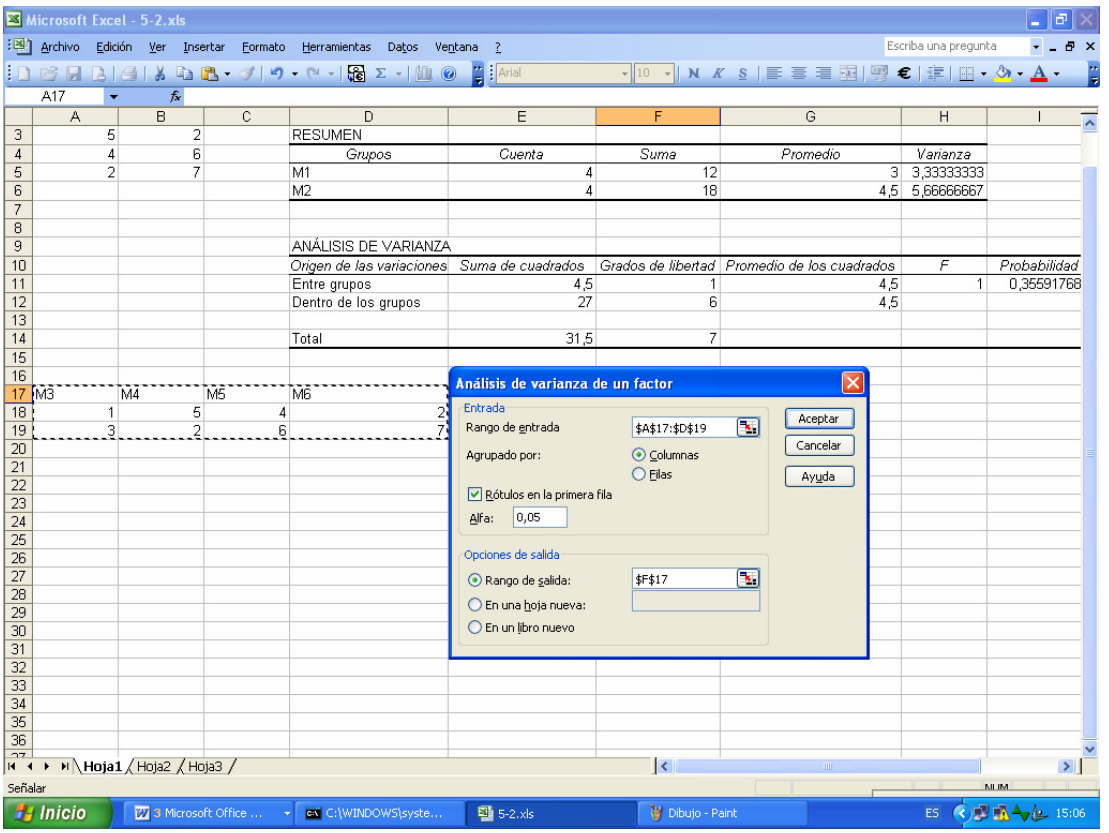


Figura 5-6

17	Análisis de varianza de un factor						
18							
19	RESUMEN						
20	*Grupos*	*Cuenta*	*Suma*	*Promedio*	*Varianza*		
21	M3	2	4	2	2		
22	M4	2	7	3,5	4,5		
23	M5	2	10	5	2		
24	M6	2	9	4,5	12,5		
25							
26							
27	ANÁLISIS DE VARIANZA						
28	*Origen de las variaciones*	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrac*	*F*	*Probabilidad*	*Valor crítico para F*
29	Entre grupos	10,5	3	3,5	0,666666667	0,615099821	6,591382117
30	Dentro de los grupos	21	4	5,25			
31							
32	Total	31,5	7				

Figura 5-7

La tabla del análisis de la varianza por estratos es entonces la siguiente:

Fuente de variación	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*
Entre estratos	$n-1=4-1=3$	$\sum_i^n\sum_j^k \left(\overline{X}_i - \overline{X}\right)^2 = 10{,}5$	$S^2_{bst} = 10{,}5/3 = 3{,}5$
Dentro de estratos	$N-n=8-4=4$	$\sum_i^n\sum_j^k \left(X_{ij} - \overline{X}_i\right)^2 = 21$	$S^2_{wst} = 21/4 = 5{,}25$
Total	$N-1=8-1=7$	$\sum_i^n\sum_j^k \left(X_{ij} - \overline{X}_j\right) = 31{,}5$	$S^2 = 31{,}5/7 = 4{,}5$

$$\sum_i^n\sum_j^k\left(\overline{X}_i-\overline{X}\right)^2 = k\sum_j^k\left(\overline{X}_i-\overline{X}\right)^2 = 2\left[(2-3{,}75)^2+(3{,}5-3{,}75)^2+(5-3{,}75)^2+(4{,}5-3{,}75)^2\right]=10{,}5$$

$$\sum_i^n\sum_j^k\left(X_{ij}-\overline{X}_i\right)^2 = (1-2)^2+(3-2)^2+(5-3{,}5)^2+(2-3{,}5)^2+(4-5)^2+(6-5)^2+(2-4{,}5)^2+(7-4{,}5)^2 = 21$$

A partir de esta equivalencia entre muestreo estratificado y muestreo sistemático podemos hallar la varianza del estimador de la media de la siguiente forma:

$$V(\hat{\overline{X}}) = V(\overline{x}_j) = (1-f)\frac{S^2_{wst}}{n} = \left(1-\frac{1}{2}\right)\frac{5{,}25}{4} = 0{,}65625$$

Se observa que ahora la varianza es ligeramente superior al caso en que no se consideraba estratificación. Ello es debido a que la selección de la unidad por estrato para la muestra no es aleatoria salvo en el primer estrato. Una medida de esa falta de aleatoriedad la proporciona el coeficiente de correlación $\rho_{\omega st}$, cuyo valor se calcula como:

$$\rho_{\omega st} = \frac{2\sum_j^k\sum_{i<z}^n\left(X_{ij}-\overline{X}_i\right)\left(X_{zj}-\overline{X}_z\right)}{n(n-1)(k-1)S^2_{wst}} = \frac{2}{4.3.1.5{,}25}\left((1-2)(5-3{,}5)+(1-2)(4-5)+\ldots+(6-5)(7-4{,}5)\right) = -0{,}047$$

El valor de $\rho_{\omega st}$ es negativo y muy pequeño, lo que indica que la falta de aleatoriedad en la selección de una unidad por estrato no es muy elevada. Para calcular el valor correcto de la varianza del estimador de la media considerando la falta de aleatoriedad se utiliza la siguiente expresión en función de $\rho_{\omega st}$:

$$V(\hat{\overline{X}}_{st}) = V(\overline{x}_{jst}) = (1-f)\frac{S^2_{wst}}{n}(1+(n-1)\rho_{\omega st}) = (1-0{,}5)\frac{5{,}25}{4}(1-(4-1)0{,}047) = 0{,}56$$

Se observa que ahora ya coincide la varianza con la calculada sin estratificar.

También podemos estimar la varianza a partir del muestreo estratificado, agrupando las 4 filas (estratos) de la población en grupos de 2, y considerando cada dos filas como un estrato del que seleccionamos dos unidades para la muestra. Tendremos:

$$\begin{array}{|cc|}\hline 1 & 3 \\ 5 & 2 \end{array} \} \textit{Estrato } 1$$
$$\begin{array}{|cc|}\hline 4 & 6 \\ 2 & 7 \end{array} \} \textit{Estrato } 2$$

$$\hat{V}(\bar{x}_1) = \frac{1-f}{n^2}\sum_{h}^{\frac{n}{2}}(x_{h1} - x_{h2})^2 = \frac{1-0{,}5}{4^2}\left[(1-5)^2 + (4-2)^2\right] = 0{,}625$$

$$\hat{V}(\bar{x}_2) = \frac{1-f}{n^2}\sum_{h}^{\frac{n}{2}}(x_{h1} - x_{h2})^2 = \frac{1-0{,}5}{4^2}\left[(3-2)^2 + (6-7)^2\right] = 0{,}0625$$

Por esta vía la menor varianza la presenta la segunda muestra.

La tabla del análisis de la varianza en el caso de estratificación es esencial en estos problemas, ya que proporciona prácticamente toda la información para realizar cálculos.

5.3. Una manzana de casas de una ciudad contiene 36 hogares numerados del 1 al 36. Los hogares con ingresos mensuales superiores a 1500 euros son los que tienen los números 3, 5-7, 11-13, 15-16, 20- 22, 25-26, 28 y 30-34.

1º) Se trata de estimar la proporción de hogares con sueldo mensual superior a 1500 euros utilizando muestreo sistemático. Comparar la precisión de una muestra sistemática 1 en 4 con una muestra aleatoria simple del mismo tamaño para estimar la proporción de hogares con sueldo mensual superior a 1500 euros. Justificar la respuesta en función del valor del coeficiente de correlación intramuestral y en función de la cuasivarianza intramuestral.

2º) Hallar el tamaño de muestra necesario para estimar la proporción de viviendas en las que los ingresos mensuales son superiores a 1500 euros para un error de muestreo de 16 centésimas. Hallar ese mismo tamaño para muestreo aleatorio simple y comentar el resultado.

Si definimos una variable dicotómica *A* a la que asignamos el valor 1 para los hogares en que los ingresos mensuales superan los 1500 euros y el valor 0 para el resto de los hogares, y clasificamos los 36 hogares en 9 filas de 4 viviendas cada una (muestreo sistemático 1 en 4) siguiendo el orden del enunciado del problema, tendremos la siguiente tabla:

0	0	1	0	1/4
1	1	1	0	3/4
0	0	1	1	1/2
1	0	1	1	3/4
0	0	0	1	1/4
1	1	0	0	1/2
1	1	0	1	3/4
0	1	1	1	3/4
1	1	0	0	1/2
5/9	5/9	5/9	5/9	5/9

Para calcular la varianza del estimador sistemático de la proporción hacemos:

$$V(\hat{P}) = \frac{1}{4}\left[\left(\frac{5}{9}-\frac{5}{9}\right)^2+\left(\frac{5}{9}-\frac{5}{9}\right)^2+\left(\frac{5}{9}-\frac{5}{9}\right)^2+\left(\frac{5}{9}-\frac{5}{9}\right)^2\right] = 0$$

También podemos calcular la varianza del estimador de la proporción como:

$$V(\hat{P}) = PQ - \frac{1}{k}\sum_{j=1}^{k}\hat{P}_j\hat{Q}_j = \frac{20}{36}\left(1-\frac{20}{36}\right) - \frac{1}{4}\left(\frac{5}{9}\frac{4}{9}+\frac{5}{9}\frac{4}{9}+\frac{5}{9}\frac{4}{9}+\frac{5}{9}\frac{4}{9}\right) = 0$$

A continuación, se construye la tabla del análisis de la varianza para la población (producción total) utilizando Excel. Como estamos clasificando los datos en 4 grupos (columnas), utilizaremos una variable *G*, que clasificará los valores de *A* (ceros o unos) por grupos (por columnas). Introducimos los valores de *G* en columnas de la hoja de cálculo de Excel y elegimos *Análisis de la varianza de un factor* en la opción *Análisis de datos* del menú *Herramientas*, rellenando su pantalla de entrada como se indica en la Figura 5-8. La Figura 5-9 presenta los resultados.

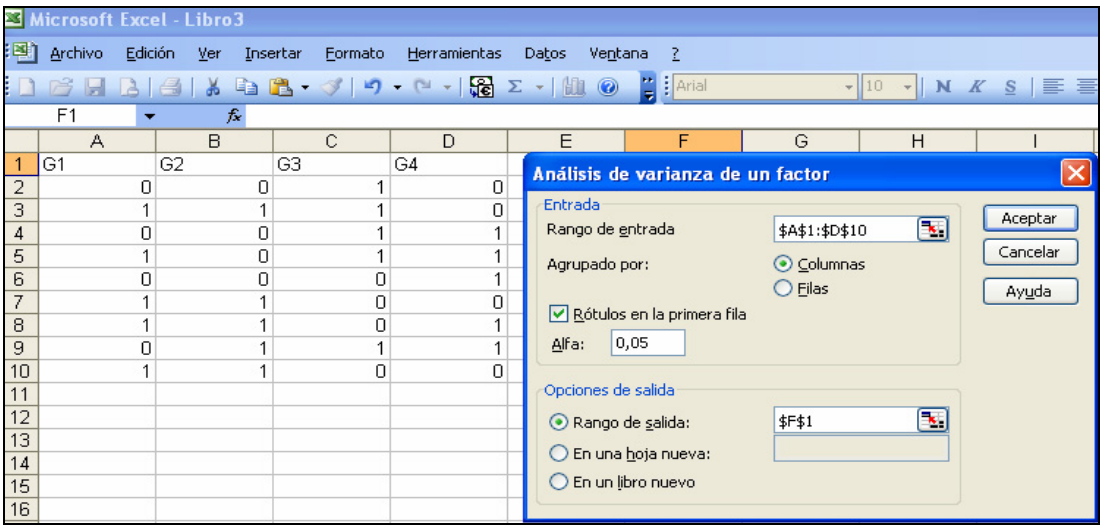


Figura 5-9

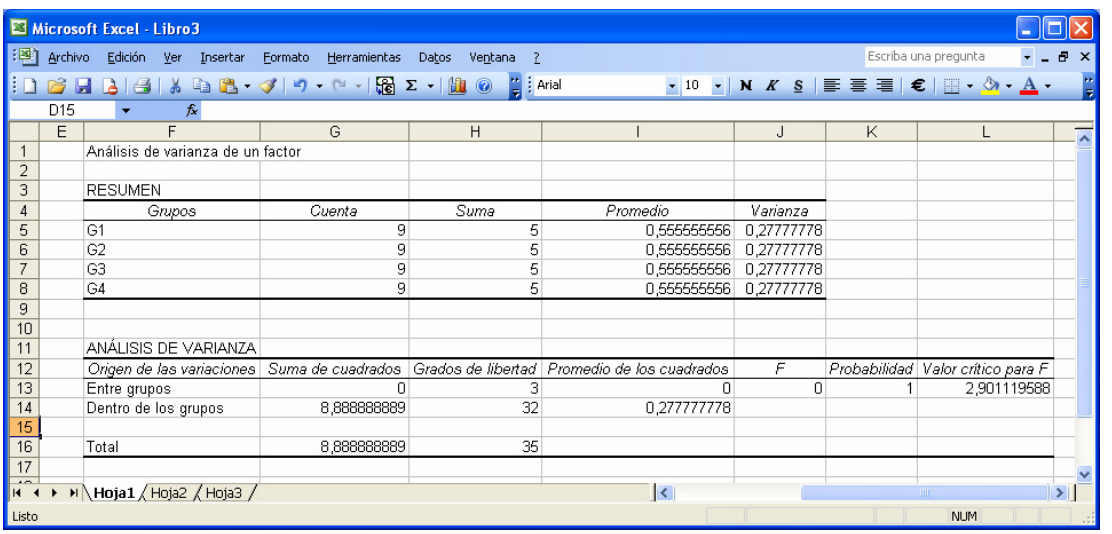


Figura 5-10

Por lo tanto, la tabla del análisis de la varianza para la población es la siguiente:

Fuente	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*
Entre	$k-1=4-1=3$	$\sum_i^n \sum_j^k \left(\bar{x}_j - \bar{X}\right)^2 = 0$	$S_{bs}^2 = 0/3 = 0$
Dentro	$N-k=36-4=32$	$\sum_i^n \sum_j^k \left(X_{ij} - \bar{x}_j\right)^2 = 8{,}88$	$S_{ws}^2 = 8{,}88/32 = 0{,}277$
Total	$N-1=36-1=35$	$\sum_i^n \sum_j^k \left(X_{ij} - \bar{X}_j\right) = 8{,}88$	$S^2 = 8{,}88/35 = 0{,}254$

Conocida esta tabla pueden realizarse ya todos los cálculos. Por ejemplo, la varianza del estimador de la proporción también podría calcularse como:

$$V(\hat{P}) = V\left(\hat{P}_j\right) = \left(1 - \frac{n}{N}\right)\frac{S_{bs}^2}{n} = \left(1 - f\right)\frac{S_{bs}^2}{n} = \left(1 - \frac{1}{4}\right)\frac{0}{9} = 0$$

Del valor de la varianza puede deducirse el valor del coeficiente de correlación intramuestral a través de la fórmula $V(\bar{x}_j) = \frac{\sigma^2}{n}(1 + (n-1)\rho_\omega)$. Tendremos:

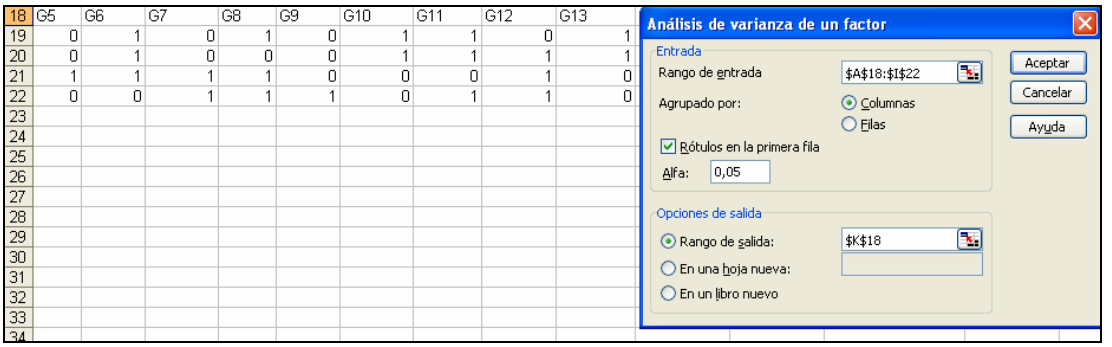
$$0 = \frac{\frac{35}{36}0{,}254}{9}(1 + (9-1)\rho_\omega) \Rightarrow \rho_\omega = -\frac{1}{8} = -\frac{1}{n-1} = -0{,}125$$

Estamos ante el caso de máxima precisión del muestreo sistemático, ya que la varianza es nula, o lo que es lo mismo, $\rho_\omega = -\frac{1}{n-1}$.

Este hecho concuerda con los valores que toman S^2 y S^2_{ws}. Concretamente $S^2 = 0{,}254 < S^2_{ws} = 0{,}277$, lo que indica que es más preciso el muestreo sistemático que el aleatorio simple. La varianza del estimador de la proporción en el muestreo aleatorio simple es (1–1/4)*0,254/9 = 0,021.

Para resolver el segundo apartado del problema consideramos ahora cada una de las 9 zonas (filas) como un estrato de 4 unidades. Tenemos entonces dividida la población en 9 estratos de 4 unidades cada uno, de modo que la muestra sistemática consta de una unidad por estrato que de forma general no es elegida aleatoriamente dentro del mismo. Esta clasificación de los elementos de la población en 9 filas de 4 unidades cada una origina una tabla del análisis de la varianza que se puede calcular con Excel.

Como estamos clasificando los datos en 9 filas (estratos), utilizaremos las variables G_5 a G_{13}, que recogen los valores de las nueve filas. A continuación elegimos *Análisis de la varianza de un factor* en la opción *Análisis de datos* del menú *Herramientas*, rellenando su pantalla de entrada como se indica en la Figura 5-11. La Figura 5-12 presenta los resultados.



	G5	G6	G7	G8	G9	G10	G11	G12	G13
19	0	1	0	1	0	1	1	0	1
20	0	1	0	0	0	1	1	1	1
21	1	1	1	1	0	0	0	1	0
22	0	0	1	1	1	0	1	1	0

Figura 5-11

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
G5	4	1	0,25	0,25
G6	4	3	0,75	0,25
G7	4	2	0,5	0,33333333
G8	4	3	0,75	0,25
G9	4	1	0,25	0,25
G10	4	2	0,5	0,33333333
G11	4	3	0,75	0,25
G12	4	3	0,75	0,25
G13	4	2	0,5	0,33333333

ANÁLISIS DE VARIANZA

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
Entre grupos	1,388888889	8	0,173611111	0,625	0,74947103	2,305313178
Dentro de los grupos	7,5	27	0,277777778			
Total	8,888888889	35				

Figura 5-12

El cuadro del análisis de la varianza por estrato es entonces el siguiente:

Fuente de variación	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*
Entre estratos	$n-1=9-1=8$	$\sum_{i}^{n}\sum_{j}^{k}\left(\overline{X}_i-\overline{X}\right)^2=1{,}388$	$S^2_{bst}=1{,}388/8=0{,}1735$
Dentro de estratos	$N-n=36-9=27$	$\sum_{i}^{n}\sum_{j}^{k}\left(X_{ij}-\overline{X}_i\right)^2=7{,}5$	$S^2_{wst}=7{,}5/27=0{,}277$
Total	$N-1=36-1=35$	$\sum_{i}^{n}\sum_{j}^{k}\left(X_{ij}-\overline{X}_j\right)=8{,}888$	$S^2=8{,}888/35=0{,}254$

Calculamos ahora el valor del coeficiente de correlación $\rho_{\omega st}$ como sigue:

$$\rho_{\omega st}=\frac{2\sum_{j}^{k}\sum_{i<z}^{n}\left(X_{ij}-\overline{X}_i\right)\left(X_{zj}-\overline{X}_z\right)}{n(n-1)(k-1)S^2_{wst}}=\frac{2}{9.8.3.0{,}277}\left((0-\frac{1}{4})(1-\frac{3}{4})+(0-\frac{1}{4})(0-\frac{1}{2})+\ldots+(1-\frac{3}{4})(0-\frac{1}{2})\right)=-0{,}125$$

Para calcular el tamaño de muestra necesario para cometer un error de muestreo igual a 0,16 despejamos n en la expresión que define la varianza de la proporción en función de $\rho_{\omega st}$. Tenemos:

$$V(\hat{P}_{st}) = (1-f)\frac{S^2_{wst}}{n}(1+(n-1)\rho_{wst}) \Rightarrow 0{,}16^2 = (1-\frac{n}{36})\frac{0{,}277}{n}(1+(n-1)(-0{,}125) \Rightarrow n = 5$$

Para calcular el tamaño de muestra anterior en muestreo aleatorio simple despejamos n en la expresión que define la varianza de la proporción en ese tipo de muestreo. Tenemos:

$$V(\hat{P}_{st}) = (1-f)\frac{S^2}{n} \Rightarrow 0{,}16^2 = (1-\frac{n}{36})\frac{0{,}254}{n} \Rightarrow n = 8$$

Obviamente el tamaño de muestra necesario para cometer el mismo error de muestreo es mayor en muestreo aleatorio simple que en muestreo sistemático, ya que en este problema el muestreo sistemático es más preciso que el muestreo aleatorio simple.

5.4. Un investigador desea determinar la calidad del azúcar contenida en la sabia de los árboles de una finca, que se encuentran situados a lo largo de la misma de forma natural en 7 hileras. El número total de árboles es desconocido, por lo que no puede realizarse una muestra irrestricta aleatoria. Como procedimiento alternativo el investigador decide usar una muestra sistemática de 1 en 7. En la tabla adjunta se encuentran los datos del contenido de azúcar en la sabia de los árboles muestreados:

Árbol muestreado	*Contenido de azúcar en la savia X*	X^2
1	82	6724
2	76	5776
3	83	6889
⋮	⋮	⋮
210	84	7056
211	80	6400
212	79	6241
	$\sum_{i=1}^{212} X_i = 17066$	$\sum_{i=1}^{212} X_i^2 = 1486800$

Estimar el contenido de azúcar promedio en la sabia de los árboles de la finca estableciendo los errores absoluto y relativo de la estimación. Realizar la estimación mediante un intervalo de confianza al nivel del 5%.

La estimación de la media vendrá dada por:

$$\hat{\bar{X}} = \bar{x}_j = \frac{\sum_{i=1}^{212} X_i}{212} = 80{,}5$$

Para calcular el error absoluto de muestreo consideramos la estimación de la varianza, que se basará en la fórmula del muestreo aleatorio simple, ya que intuitivamente podemos suponer que la población de árboles en la finca es aleatoria en cuanto al contenido de azúcar en la sabia debido a que suponemos una distribución natural de los mismos en la finca. Previamente necesitamos estimar la cuasivarianza mediante:

$$\hat{S}^2 = \bar{x}_j = \frac{\sum_{i=1}^{212} X_i^2 - \left(\sum_{i=1}^{212} X_i\right)^2 \Big/ 212}{212-1} = 535{,}48$$

Además, al ser la muestra sistemática 1 en 7 y $n = 212$ entonces $N = nk = 212.7 = 1484$ árboles. La estimación de la varianza del estimador de la media será:

$$\hat{V}(\hat{\bar{X}}) = \hat{V}(\bar{x}_j) = \left(1 - \frac{1}{7}\right)\frac{535{,}48}{212} = 2{,}16 \Rightarrow \hat{\sigma}(\hat{\bar{X}}) = 1{,}47$$

El error relativo de muestreo será:

$$\hat{C}v(\hat{\bar{X}}) = \frac{\sqrt{\hat{V}(\bar{x}_j)}}{E(\bar{x}_j)} = \frac{1{,}47}{\hat{\bar{X}}} = \frac{1{,}47}{80{,}5} = 0{,}0182 \quad (1{,}82\%)$$

El error relativo es bajo, por lo que la estimación puede ser buena. Por otra parte, un intervalo de confianza para la media suponiendo normalidad en la población será:

$$\hat{\bar{X}} \pm \lambda_\alpha \hat{\sigma}(\hat{\bar{X}}) = 80{,}5 \pm 1{,}96 \cdot 1.47 = [77{,}6\ 83{,}4]$$

En caso de no poder suponer normalidad se toma el intervalo más tosco dado por:

$$\hat{\bar{X}} \pm \frac{\hat{\sigma}(\hat{\bar{X}})}{\sqrt{\alpha}} = 80{,}5 \pm \frac{1.47}{\sqrt{0{,}05}} = [74,\ 87]$$

El intervalo para no normalidad es más ancho (peor) que en el caso de normalidad, pero no demasiado.

5.5. Un hortelano tiene un huerto experimental con $N = 1300$ manzanos de una nueva variedad en estudio. El investigador desea estimar la producción total (en quintales) de la huerta, con base en los manzanos de una muestra sistemática de 1 en 10. La media y la varianza muestrales para los árboles muestreados fueron $\bar{x}_j = 3{,}52$ quintales y $\hat{S}^2 = 0{,}48$ quintales. Utilizar estos datos para estimar la producción total, y establecer un límite para el error de estimación.

La estimación de la producción total estará dada por:

$$\hat{X} = N\bar{x}_j = 1300(3{,}52) = 4576 \text{ quintales}$$

Para calcular el error absoluto de muestreo consideramos la estimación de la varianza, que se basará en la fórmula del muestreo aleatorio simple, ya que intuitivamente podemos suponer que la población de manzanos en el huerto es aleatoria debido a que suponemos una distribución natural de los mismos en el huerto.

Además, al ser la muestra sistemática 1 en 10 y $N = 1300$ entonces $N = nk \Rightarrow 1300 = n.10 \Rightarrow n = 130$ manzanos árboles. La estimación de la varianza del estimador de la media será:

$$\hat{V}(\hat{X}) = N^2 \hat{V}(\bar{x}_j) = 1300^2 \left(1 - \frac{130}{1300}\right) \frac{0{,}48}{130} = 5625 \Rightarrow \hat{\sigma}(\hat{X}) = 75$$

El error relativo de muestreo será:

$$\hat{C}v(\hat{X}) = \frac{\sqrt{\hat{V}(\hat{X})}}{\hat{X}} = \frac{75}{4576} = \frac{1{,}47}{80{,}5} = 0{,}016 \quad (1{,}6\%)$$

El error relativo es bajo, por lo que la estimación puede ser buena. Por otra parte, un intervalo de confianza al 95% para la producción total suponiendo normalidad en la población será:

$$\hat{X} \pm \lambda_\alpha \hat{\sigma}(\hat{X}) = 4576 \pm 2 \cdot 75 = [4426 \;\; 4726]$$

El límite para el error de estimación está dado por:

$$2\hat{\sigma}(\hat{X}) = 150$$

5.6. Una muestra sistemática de 1 en 10 es obtenida de una lista de votantes registrados para estimar la proporción de votantes que están a favor de la emisión de bonos propuesta. Se utilizan diferentes puntos de inicio aleatorio para asegurar que los resultados de la muestra no se ven afectados por variación periódica en la población. Los resultados codificados de esta encuesta de elección previa se muestran en la tabla adjunta. Estimar p, la proporción de los 5775 votantes registrados que están a favor de la emisión de bonos propuesta ($N = 5775$). Establecer un límite para el error de estimación.

Votante	Respuesta
4	1
10	0
16	1
.	.
.	.
.	.
5760	0
5766	0
5772	1

$$\sum_{i=1}^{962} y_i = 652$$

Al ser la muestra sistemática 1 en 6 y $N = 5775$ entonces $N = nk \Rightarrow 5775 = n*6 \Rightarrow E(n) = 962$ donde $E(n)$ significa parte entera de n. Por tanto, el tamaño muestral es 962.

Como n es grande y se han tomado varios puntos de inicio aleatorio en la extracción de la muestra sistemática, podemos estimar la proporción proporcional mediante la proporción muestral, y el error se estimará utilizando la fórmula del muestreo aleatorio simple. Tenemos:

$$\hat{P} = \hat{P}_j = \frac{\sum_{i=1}^{212} X_i}{962} = \frac{652}{962} = 0{,}678$$

$$\hat{V}(\hat{P}) = \hat{V}(\hat{P}_j) = \left(1 - \frac{n}{N}\right)\frac{\hat{P}_j(1-\hat{P}_j)}{n-1} = \left(1 - \frac{962}{5775}\right)\frac{0{,}678(1-0{,}678)}{962-1} = 0{,}000196 \Rightarrow \hat{\sigma}(\hat{\bar{X}}) = 0{,}014$$

El error relativo de muestreo cuando se asegura que el 67,8% de los votantes registrados favorece la emisión de bonos propuesta, será:

$$\hat{C}v(\hat{P}) = \frac{\sqrt{\hat{V}(\hat{P})}}{\hat{P}} = \frac{0{,}014}{0{,}678} = \frac{1{,}47}{80{,}5} = 0{,}0206 \quad (2{,}06\%)$$

Por otra parte, un intervalo de confianza para la proporción, suponiendo normalidad en la población será:

$$\hat{P} \pm \lambda_\alpha \hat{\sigma}(\hat{P}) = 0{,}678 \pm 2 \cdot 0{,}014$$

El límite para el error de estimación será el radio del intervalo de confianza, o sea, 0,028 (2,8%).

5.7. Un parque estatal cobra la admisión por automóvil en lugar de por persona, y un funcionario del parque quiere estimar el número promedio de personas por automóvil para un día concreto en particular durante el verano. El funcionario sabe por experiencia que entrarán al parque alrededor de 400 automóviles y quiere muestrear 80 de ellos. Para obtener una estimación de la varianza, utiliza el muestreo sistemático replicado con 10 muestras de 8 automóviles cada una. En la tabla siguiente se presentan los datos del número de personas por automóvil (entre paréntesis):

Punto de inicio aleatorio	Segundo elemento	Tercer elemento	Cuarto elemento	Quinto elemento	Sexto elemento	Séptimo elemento	Octavo elemento	$\bar{y}_i$
2 (3)	52 (4)	102 (5)	152 (3)	202 (69	252 (1)	302 (4)	352 (4)	3,75
5 (5)	55 (3)	105 (4)	155 (2)	205 (4)	255 (2)	305 (3)	355 (4)	3,38
7 (2)	57 (4)	107 (6)	157 (2)	207 (3)	257 (2)	307 (1)	357 (3)	2,88
13 (6)	63 (4)	113 (6)	163 (7)	213 (2)	263 (3)	313 (2)	363 (7)	4,62
26 (4)	76 (5)	126 (7)	176 (4)	226 (2)	276 (6)	326 (2)	376 (6)	4,5
31 (7)	81 (6)	131 (4)	181 (4)	231 (3)	281 (6)	331 (7)	381 (5)	5,25
35 (3)	85 (3)	135 (2)	185 (3)	235 (6)	285 (5)	335 (6)	385 (8)	4,5
40 (2)	90 (6)	140 (2)	190 (5)	240 (5)	290 (4)	340 (4)	390 (5)	4,12
45 (2)	95 (6)	145 (3)	195 (6)	245 (4)	295 (4)	345 (5)	395 (4)	4,25
46(6)	96 (5)	146 (4)	196 (6)	246 (3)	296 (3)	346 (5)	396 (3)	4,38

Estimar el número promedio de personas por automóvil y establezcer un límite para el error de estimación.

Como tenemos varios arranques aleatorios, utilizaremos el método de las muestras interpenetrantes.

Podemos formar un estimador combinado de la media poblacional basado en las medias de las *t* muestras (cada media muestral es un estimador insesgado de la misma media poblacional) promediando las medias de las 10 muestras sistemáticas (filas de la tabla del enunciado) de la siguiente forma:

$$\bar{x}_c = \frac{1}{t}\sum_{1}^{t} \bar{x}_t = \frac{1}{10}(3{,}75 + 3{,}38 + \cdots + 4{,}38) = 4{,}16$$

El estimador insesgado de su varianza mediante la aplicación del método de las muestras interpenetrantes es:

$$\hat{V}(\bar{x}_c) = (1 - n/N)\frac{1}{t(t-1)}\sum_{i}^{t} \bar{x}_i^2 - t\bar{x}_c^2 = (1 - n/N)\frac{1}{t(t-1)}\left(\sum_{i}^{t} \bar{x}_i^2 - \sum_{i}^{t} \bar{x}_c^2\right) = (1 - n/N)\frac{1}{t(t-1)}\sum_{i}^{t}\left(\bar{x}_i^2 - \bar{x}_c^2\right)$$

$$\hat{V}(\bar{x}_c) = (1 - 80/400)\frac{1}{10(9-1)}177410 - 10 * 4{,}16^2 = 0{,}0365$$

El límite para el error de estimación al 95% es $2\sqrt{\hat{V}(\bar{x}_c)} = 2\sqrt{0{,}0365} = 0{,}38$.

5.8. Una empresa publicitaria está iniciando una campaña de promoción para un nuevo producto. La empresa quiere muestrear clientes potenciales en una pequeña comunidad para determinar la aceptación del producto. Para eliminar algo de los costos asociados con las entrevistas personales, el investigador decide seleccionar una muestra sistemática de entre *N* = 5000 nombres listados en un registro de la comunidad y recolectar los datos mediante entrevistas por teléfono. Determinar el tamaño de muestra requerido para estimar la proporción de personas que consideran <<aceptable>> el producto, con un límite para el error de estimación de magnitud 0,03 (esto es, 3%).

Como el límite para el error de la estimación es 0,003, tenemos:

$$2\sqrt{\hat{V}(\hat{P})} = 0{,}03 \Rightarrow \hat{V}(\hat{P}) = 0{,}000225$$

Entonces, el tamaño de muestra requerido es:

$$n = \frac{N\hat{P}\hat{Q}}{(N-1)\hat{V}(\hat{P}) + \hat{P}\hat{Q}} = \frac{5000(0{,}5)(0.5)}{4999(0{,}000225) + (0{,}5)(0{,}5)} = 909{,}240 \approx 910$$

La empresa debe entrevistar a 910 personas para determinar la aceptación del producto, con un límite para el error de estimación del 3%.

Se ha supuesto que es correcta la aproximación del error del muestreo sistemático por el error del muestreo aleatorio simple (población grande de carácter típicamente aleatorio) y se ha tomado $\hat{P} = \hat{Q} = 1/2$ por desconocimiento de sus valores.

EJERCICIOS PROPUESTOS

5.1. Dada la población siguiente:

u_i	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9
X_i	1	3	5	2	4	6	2	7	3

se desea obtener una muestra sistemática de tamaño 3 (1 en 3). Determinar el espacio muestral y las probabilidades asociadas a las muestras posibles para este tipo de muestreo. Calcular las varianzas de los estimadores insesgados del total y de la media. Estimar dichas varianzas y comparar la precisión de este tipo de muestreo con la del muestreo aleatorio simple. Seleccionar la muestra más precisa.

5.2. En un directorio de 13 casas de una calle las personas están distribuidas hogar a hogar como sigue:

1	2	3	4	5	6	7	8	9	10	11	12	13
M	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*
F	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*	*F*
f	*f*	*m*		*m*	*f*	*f*	*m*	*m*	*m*	*f*	*f*	
m	*m*	*f*		*m*	*m*	*f*	*f*		*f*	*m*		
f	*f*			*f*		*m*						

M=varón adulto, F=mujer adulta, m=hijo varón, f=hija

Se realiza muestreo sistemático de una de cada 5 personas (muestreo 1 en 5), numerando los elementos de la población por columnas hacia abajo y luego yendo a la parte superior de la siguiente columna (se empieza por la primera columna de la izquierda). Se pide lo siguiente:

1º) Calcular el valor del coeficiente de correlación $\rho_{\omega st}$ y hallar la varianza del estimador de la proporción de varones adultos en la población utilizando la relación entre muestreo sistemático y muestreo estratificado.

2º) ¿Qué muestra sistemática es la mejor? ¿Cuál es la proporción estimada de varones adultos en la población?

5.3. La administración de una empresa de servicio público está interesada en la cantidad promedio de tiempo que llevan vencidas las cuentas atrasadas. Una muestra sistemática será extraída de una lista en orden alfabético con $N = 2500$ cuentas de clientes que están vencidas. En una encuesta similar realizada el año anterior, la varianza muestral fue $s^2 = 100$ días. Determinar el tamaño de muestra requerido para estimar μ, la cantidad promedio de tiempo que tienen de estar vencidas las cuentas de la empresa de servicio público, con un límite para el error de estimación de 2 días.

CAPÍTULO

6

MUESTREO POR MÉTODOS INDIRECTOS. RAZÓN, REGRESIÓN Y DIFERENCIA

OBJETIVOS

1. Presentar el concepto de estimación no lineal.

2. Presentar el concepto de estimación por métodos indirectos.

3. Analizar el estimador de razón, su sesgo y su varianza.

4. Estimar el sesgo y la varianza del estimador por razón.

5. Analizar los estimadores y sus errores en muestreo estratificado con reposición.

6. Comprender la formación de estimadores de magnitudes poblacionales basados en la razón.

7. Estudiar los errores y su estimación para estimadores indirectos basados en la razón.

8. Analizar el estimador de razón, su sesgo y su varianza.

9. Obtener la varianza mínima para el estimador de regresión y su estimación.

10. Comparar la estimación indirecta por regresión con otros tipos de muestreo.

11. Analizar el estimador por diferencia, sesgo, varianza y sus estimaciones.

12. Comprender los métodos indirectos en muestreo estratificado.

13. Analizar la estimación por razón en muestreo estratificado.

14. Analizar la estimación por regresión en muestreo estratificado.

15. Diferenciar entre estimadores separados y estimadores combinados.

16. Comparar las precisiones de los métodos de estimación indirecta con estratificación.

ÍNDICE

1. Estimadores no lineales.

2. Muestreo por métodos indirectos. El estimador de razón.

3. Estimaciones de los parámetros poblacionales basadas en la razón y errores.

4. Estimaciones por regresión y errores.

5. Estimaciones por diferencia y errores.

6. Estimadores de razón en el muestreo estratificado.

7. Estimadores de regresión en el muestreo estratificado.

8. Problemas resueltos.

9. Ejercicios propuestos.

ESTIMADORES NO LINEALES

Al estimar un parámetro poblacional la dificultad principal suele estar en el cálculo del error de muestreo (raíz cuadrada de la varianza del estimador). Por esta razón, son muchos los procedimientos analizados para la estimación de varianzas. Según Wotter (1985), podemos clasificar las situaciones que se pueden presentar atendiendo a la naturaleza del parámetro a estimar (parámetros lineales o no lineales) y al diseño muestral utilizado (diseños simples o complejos). Se pueden estimar parámetros lineales en diseños simples, parámetros no lineales en diseños simples, parámetros lineales en diseños complejos o parámetros no lineales en diseños complejos. Aunque la mayor parte de la teoría básica de muestreo se basa en el cálculo de estimadores de parámetros lineales en diseños simples, también se han desarrollado procedimientos para aproximación lineal de estimadores que están basados en un desarrollo en serie de Taylor para obtener una aproximación lineal del estimador y así poder aplicar posteriormente toda la teoría desarrollada para estimadores lineales.

Otras técnicas, como los métodos de replicación de muestras, se basan en la generación de diversas muestras, todas bajo el mismo diseño muestral, con el fin de obtener información acerca de la distribución del estimador.

Además existen otras técnicas, como los métodos de exploración intensiva de una muestra, que consisten en la generación de muestras a partir de la muestra inicial, obtenida mediante un determinado diseño, usando técnicas muy variadas.

Estimadores no lineales. Método general de linealización para la estimación de varianzas

Supongamos un parámetro poblacional θ del cual hemos obtenido un estimador no lineal $\hat{\theta} = f(x_1,...,x_n)$ basado en la muestra $(x_1,...,x_n)$. Se trata de expresar dicho estimador como función de una serie de estimadores $\hat{\theta}_1,...,\hat{\theta}_k$, es decir, $\hat{\theta} = f(x_1,...,x_n) = \varphi(\hat{\theta}_1,...,\hat{\theta}_k)$, de modo que si se calculan las varianzas de los nuevos estimadores habremos conseguido resolver nuestro problema.

Supongamos que $\hat{\theta}_1,...,\hat{\theta}_k$ son estimadores insesgados de $\theta_1,...,\theta_k$ respectivamente y que los valores teóricos cumplen $\theta = \varphi(\theta_1,...,\theta_k)$. El desarrollo de Taylor de $\varphi(\hat{\theta}_1,...,\hat{\theta}_k)$ en un entorno del punto $(\theta_1,...,\theta_k)$ es el siguiente:

$$\varphi(\hat{\theta}_1,...,\hat{\theta}_k) = \varphi(\theta_1,...,\theta_k) + d\varphi(\hat{\theta}_1,...,\hat{\theta}_k)_{(\theta_1,...,\theta_k)} + T_n$$

donde T_n es el término complementario o resto, el cual puede ser despreciado o no dependiendo de las condiciones del entorno. Para un entorno suficientemente pequeño supongamos que T_n sí es despreciable, resultando:

$$\hat{\theta} - \theta \approx d\varphi(\hat{\theta}_1,...,\hat{\theta}_k)_{(\theta_1,...,\theta_k)} = \sum_{r=1}^{k} \left(\frac{\partial \varphi(\hat{\theta}_1,...,\hat{\theta}_k)}{\partial \hat{\theta}_r} \right)_{(\theta_1,...,\theta_k)} (\hat{\theta}_r - \theta_r)$$

Elevando ambos términos de esta igualdad al cuadrado y tomando esperanzas obtenemos una expresión aproximada para la varianza de $\hat{\theta}$, es decir,

$$V\left(\hat{\theta}\right)=E\left(\hat{\theta}-\theta\right)^2\approx E\left[\sum_{r=1}^{k}\left(\frac{\partial\varphi\left(\hat{\theta}_1,...,\hat{\theta}_k\right)}{\partial\hat{\theta}_r}\right)_{(\theta_1,...,\theta_k)}\left(\hat{\theta}_r-\theta_r\right)\right]^2$$

$$=E\left[\sum_{r=1}^{k}\sum_{l=1}^{k}\left(\frac{\partial\varphi\left(\hat{\theta}_1,...,\hat{\theta}_k\right)}{\partial\hat{\theta}_r}\right)_{(\theta_1,...,\theta_k)}\left(\hat{\theta}_r-\theta_r\right)\left(\frac{\partial\varphi\left(\hat{\theta}_1,...,\hat{\theta}_k\right)}{\partial\hat{\theta}_l}\right)_{(\theta_1,...,\theta_k)}\left(\hat{\theta}_l-\theta_l\right)\right]$$

$$=\sum_{r=1}^{k}\sum_{l=1}^{k}\left(\frac{\partial\varphi\left(\hat{\theta}_1,...,\hat{\theta}_k\right)}{\partial\hat{\theta}_r}\right)_{(\theta_1,...,\theta_k)}\left(\frac{\partial\varphi\left(\hat{\theta}_1,...,\hat{\theta}_k\right)}{\partial\hat{\theta}_l}\right)_{(\theta_1,...,\theta_k)}Cov\left(\hat{\theta}_r,\hat{\theta}_l\right)$$

Aplicación al cociente de estimadores

Sea $R=\dfrac{\alpha}{\beta}$ un parámetro poblacional y $\hat{R}=\dfrac{\hat{\alpha}}{\hat{\beta}}$ un estimador del mismo.

Observamos que $\hat{R}=\varphi(\hat{\alpha},\hat{\beta})$ y $R=\varphi(\alpha,\beta)$, por lo que estamos en condiciones de aplicar el método general de linealización de varianzas previamente explicado. Haciendo un desarrollo en serie de Taylor de la función $\varphi(\hat{\alpha},\hat{\beta})$ en el punto (α,β) resulta:

$$\hat{R}-R\approx\left(\frac{\partial\varphi\left(\hat{\alpha},\hat{\beta}\right)}{\partial\hat{\alpha}}\right)_{(\alpha,\beta)}\left(\hat{\alpha}-\alpha\right)+\left(\frac{\partial\varphi\left(\hat{\alpha},\hat{\beta}\right)}{\partial\hat{\beta}}\right)_{(\alpha,\beta)}\left(\hat{\beta}-\beta\right)$$

y elevando al cuadrado y tomando esperanzas tenemos:

$$V\left(\hat{R}\right)=E\left(\hat{R}-R\right)^2\approx E\left[\left(\frac{\partial\varphi\left(\hat{\alpha},\hat{\beta}\right)}{\partial\hat{\alpha}}\right)_{(\alpha,\beta)}\left(\hat{\alpha}-\alpha\right)+\left(\frac{\partial\varphi\left(\hat{\alpha},\hat{\beta}\right)}{\partial\hat{\beta}}\right)_{(\alpha,\beta)}\left(\hat{\beta}-\beta\right)\right]^2$$

$$=\left(\frac{\partial\hat{R}}{\partial\hat{\alpha}}\right)^2_{(\alpha,\beta)}V\left(\hat{\alpha}\right)+\left(\frac{\partial\hat{R}}{\partial\hat{\beta}}\right)^2_{(\alpha,\beta)}V\left(\hat{\beta}\right)+2\left(\frac{\partial\hat{R}}{\partial\hat{\alpha}}\right)_{(\alpha,\beta)}\left(\frac{\partial\hat{R}}{\partial\hat{\beta}}\right)_{(\alpha,\beta)}Cov\left(\hat{\alpha},\hat{\beta}\right)$$

$$=\frac{1}{\beta^2}V\left(\hat{\alpha}\right)+\left(-\frac{\alpha}{\beta^2}\right)^2V\left(\hat{\beta}\right)+2\frac{1}{\beta}\left(-\frac{\alpha}{\beta^2}\right)Cov\left(\hat{\alpha},\hat{\beta}\right)$$

$$=\frac{1}{\beta^2}\left[V\left(\hat{\alpha}\right)+R^2V\left(\hat{\beta}\right)-2RCov\left(\hat{\alpha},\hat{\beta}\right)\right]$$

Otra expresión alternativa para la varianza de $\hat{R}=\varphi(\hat{\alpha},\hat{\beta})$ es:

$$V\left(\hat{R}\right)\approx R^2\left[\frac{V\left(\hat{\alpha}\right)}{\alpha^2}+\frac{V\left(\hat{\beta}\right)}{\beta^2}-2\frac{Cov\left(\hat{\alpha},\hat{\beta}\right)}{\alpha\beta}\right]$$

MUESTREO POR MÉTODOS INDIRECTOS. EL ESTIMADOR DE RAZÓN

Los métodos indirectos utilizan la información conocida relativa a una variable auxiliar Y (variable de apoyo) correlacionada con la variable en estudio X para conseguir estimaciones más precisas para X que las calculadas únicamente a partir de la muestra de la variable que se estudia.

Entre los métodos clásicos de estimación indirecta más utilizados se encuentran el método de *estimación por razón* (basado en la razón entre X e Y), el método de *estimación por regresión* (basado en la regresión entre X e Y) y el método de *estimación por diferencia* (basado en la diferencia entre X e Y). Estos tres métodos serán desarrollados a lo largo de este capítulo.

La estimación indirecta constituye el complemento de la estimación directa. No se trata por sí solo de un método eficiente de estimación, pero junto con la estimación directa desarrolla casi totalmente la información muestral. Los métodos de estimación indirecta aprovechan la información de variables auxiliares correlacionadas con la variable objeto de estudio con el fin de conseguir una ganancia en precisión de los estimadores.

Sea X la variable objetivo y supongamos que se conoce $Y = \sum_{i=1}^{N} Y_i$, donde (X_i, Y_i) se corresponden con los pares de valores de las variables X e Y respectivamente, observados en la unidad i-ésima de la población o de la muestra. Nuestro objetivo es obtener un estimador para X que sea más preciso que el estimador directo basado únicamente en la muestra. La expresión general de los estimadores indirectos es la siguiente:

$$f\left(\hat{X}_G\right) = f\left(\hat{X}\right) + b_0\left(f(Y) - f\left(\hat{Y}\right)\right)$$

siendo f una función, $\hat{X}_G$ el estimador indirecto de X, $\hat{X}$ e $\hat{Y}$ los estimadores directos de X e Y, respectivamente, y b_0 un coeficiente de corrección que, dependiendo de su valor, nos dará los diferentes tipos de estimadores indirectos. Como caso particular supongamos $f(x) = x$. Entonces $\hat{X}_G = \hat{X} + b_0\left(Y - \hat{Y}\right)$.

Los casos más frecuentes de estimadores indirectos son los siguientes:

1. Si $b_0 = 0$, se tiene $\hat{X}_G = \hat{X}$, es decir, el estimador obtenido es el directo.
2. Si $b_0 = 1$, entonces $\hat{X}_G = \hat{X} + \left(Y - \hat{Y}\right)$, denominado estimador de la diferencia o diferencial.
3. Si $b_0 = \dfrac{\hat{X}}{\hat{Y}} = \hat{R}$, se obtiene el estimador de razón.

$$\hat{X}_G = \hat{X} + \frac{\hat{X}}{\hat{Y}}\left[Y - \hat{Y}\right] = \frac{\hat{X}}{\hat{Y}}Y = \hat{R}Y = \hat{X}_R$$

4. Si $b_0 = b$, se obtiene el estimador de regresión.

$$\hat{X}_G = \hat{X} + b\left(Y - \hat{Y}\right) = \hat{X}_{rg}$$

Supongamos una población formada por N unidades, $\{U_1,...,U_N\}$, y nos fijamos en dos características (X,Y) para cada unidad, siendo X la variable objeto de estudio e Y una variable auxiliar correlacionada con X. Llamaremos *razón* a $R=\frac{X}{Y}$ y su estimador viene dado por la expresión:

$$\hat{R}=\frac{\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}y_i}=\frac{\hat{X}}{\hat{Y}}=\frac{\bar{x}}{\bar{y}}$$

A partir de la razón podemos también estimar totales y medias mediante:

$$\hat{X}_R=\hat{R}Y$$
$$\hat{\bar{X}}_R=\hat{R}\bar{Y}$$

Estos estimadores no son insesgados pero tienen varianza muy pequeña y otras propiedades que los hacen deseables. Sin embargo, es preciso conocer Y o $\bar{Y}$ para poder calcularlos.

$\hat{R}$ es consistente, pero en general es sesgado. Para muestras grandes, $\hat{R}\rightarrow N(R,V(\hat{R}))$ y el sesgo es despreciable. No se conoce la expresión exacta de la varianza de $\hat{R}$, aunque bajo ciertas condiciones se puede obtener una expresión aproximada de la misma. Podemos expresar el sesgo en función del coeficiente de correlación entre $\hat{R}$ e $\bar{y}$ del siguiente modo:

$$B(\hat{R})=-\frac{Cov(\hat{R},\bar{y})}{\bar{Y}}=-\frac{\rho\sigma_{\hat{R}}\sigma_{\bar{y}}}{\bar{Y}}$$

$\left|\frac{B(\hat{R})}{\sigma_{\hat{R}}}\right|$ es una medida del sesgo por unidad de desviación típica, es decir, una medida relativa del sesgo respecto del error de muestreo. Además, si $\left|\frac{B(\hat{R})}{\sigma_{\hat{R}}}\right|$ es del orden del 10%, entonces el sesgo puede ser considerado despreciable en relación al error estándar.

Se cumple que $B(\hat{R})=0\Leftrightarrow \hat{R}$ e $\bar{y}$ *son variables incorreladas en el muestreo*, con lo que *ya tenemos la primera de las condiciones para la insesgadez del estimador de la razón*. Además se cumple que:

$$B\left(\hat{R}\right)=-\rho_{(\hat{R},\bar{y})}\sigma_{\hat{R}}Cv(\bar{y})\Rightarrow\left|\frac{B\left(\hat{R}\right)}{\sigma_{\hat{R}}}\right|=\left|\rho_{(\hat{R},\bar{y})}\right|\cdot Cv(\bar{y})\leq Cv(\bar{y})$$

con lo que el sesgo relativo (módulo del cociente entre el sesgo del estimador de la razón y su desviación típica) está acotado por el coeficiente de variación de $\bar{y}$.

Entonces, *para que el sesgo del estimador de la razón sea despreciable* bastará con que el coeficiente de variación de la media muestral de la variable auxiliar sea menor que 1/10, ya que en este caso:

$$\left|\frac{B(\hat{R})}{\sigma_{\hat{R}}}\right| \leq Cv(\bar{y}) < \frac{1}{10}$$

Se observa que el sesgo relativo es tanto menor cuanto menor sea $Cv(\bar{y})$. Además, para intentar eliminar la influencia del sesgo se tomarán tamaños de muestra tales que el sesgo sea despreciable, es decir, tamaños de muestra tales que $Cv(\bar{y}) < 1/10$. Para hallar este tamaño de muestra en el muestreo sin reposición operamos como se indica a continuación:

$$Cv(\bar{y}) = \frac{\sigma(\bar{y})}{E(\bar{y})} = \frac{\sqrt{V(\bar{y})}}{\bar{Y}} = \frac{\sqrt{\left(1-\frac{n}{N}\right)\frac{S_Y^2}{n}}}{\bar{Y}} < \frac{1}{10} \Rightarrow n > \frac{100NS_Y^2}{N\bar{y}^2 + 100S_Y^2} = \frac{100N\frac{S_Y^2}{\bar{y}^2}}{N + 100\frac{S_Y^2}{\bar{y}^2}}$$

Para hallar el tamaño de muestra para el que el sesgo es despreciable en el muestreo con reposición operamos como se indica a continuación:

$$Cv(\bar{y}) = \frac{\sigma(\bar{y})}{E(\bar{y})} = \frac{\sqrt{V(\bar{y})}}{\bar{Y}} = \frac{\sqrt{\frac{\sigma_Y^2}{n}}}{\bar{Y}} < \frac{1}{10} \Rightarrow n > \frac{100\sigma_Y^2}{\bar{Y}^2} = 100\frac{\sigma_Y^2}{\bar{Y}^2}$$

La segunda condición de insesgadez del estimador de la razón es que *si la recta de regresión de la variable auxiliar Y sobre la variable en estudio X (o la de X sobre Y) pasa por el origen de coordenadas entonces el estimador de la razón* $\hat{R}$ *es insesgado para R*.

Cálculo aproximado del sesgo del estimador de razón y su estimación

El sesgo del estimador de razón puede aproximarse como sigue:

Muestreo sin reposición

$$B(\hat{R}) = \frac{(1-f)}{n\bar{Y}^2}\left(RS_Y^2 - S_{XY}\right)$$

Muestreo con reposición

$$B(\hat{R}) = \frac{1}{n\bar{Y}^2}\left(R\sigma_Y^2 - \sigma_{XY}\right)$$

Estimación del sesgo del estimador de la razón

La expresión obtenida para el sesgo del estimador de la razón va a permitir se estimación a partir de los valores muestrales:

Muestreo sin reposición

Como en muestreo sin reposición las cuasivarianzas poblacionales se estiman insesgadamente por cuasivarianzas muestrales, tenemos:

$$\hat{B}(\hat{R}) = \frac{(1-f)}{n\overline{Y}^2}\left(\hat{R}\hat{S}_Y^2 - \hat{S}_{XY}\right)$$

Muestreo con reposición

Como en muestreo con reposición las varianzas poblacionales se estiman insesgadamente por cuasivarianzas muestrales, tenemos:

$$\hat{B}(\hat{R}) = \frac{1}{n\overline{Y}^2}\left(\hat{R}\hat{S}_Y^2 - \hat{S}_{XY}\right)$$

Varianza aproximada del estimador de la razón

Muestreo sin reposición

$$V\left(\hat{R}\right) = \frac{1-f}{\overline{Y}^2 n}\cdot\left(S_x^2 + R^2 S_y^2 - 2RS_{xy}\right) = \frac{1-f}{\overline{Y}^2 n(N-1)}\cdot\left[\sum_i^N X_i^2 + R^2\sum_i^N Y_i^2 - 2R\sum_i^N X_i Y_i\right]$$

Muestreo con reposición

$$V\left(\hat{R}\right) = \frac{1}{\overline{Y}^2 n}\cdot\left(\sigma_x^2 + R^2\sigma_y^2 - 2R\sigma_{xy}\right) = \frac{1}{\overline{Y}^2 nN}\cdot\left[\sum_i^N X_i^2 + R^2\sum_i^N Y_i^2 - 2R\sum_i^N X_i Y_i\right]$$

Estimación de la varianza del estimador de la razón

Muestreo sin reposición

Utilizaremos que las cuasivarianzas muestrales estiman insesgadamente las cuasivarianzas poblacionales ($\hat{S}_x^2$ estimador insesgado de S^2_x, $\hat{S}_{Yx}^2$ estimador insesgado de S^2_y y $\hat{S}_{XY}$ estimador insesgado de S_{xy}). A su vez, utilizaremos el estimador reciente obtenido para la razón R. Tenemos:

$$\hat{V}\left(\hat{R}\right) = \frac{1-f}{\overline{Y}^2 n}\cdot\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right) = \frac{1-f}{\overline{Y}^2 n(n-1)}\cdot\left[\sum_i^n X_i^2 + \hat{R}^2\sum_i^n Y_i^2 - 2\hat{R}\sum_i^n X_i Y_i\right]$$

Muestreo con reposición

Utilizaremos el hecho de que las cuasivarianzas muestrales estiman insesgadamente las varianzas poblacionales ($\hat{S}_x^2$ estimador insesgado de σ^2_x, $\hat{S}_{Yx}^2$ estimador insesgado de σ^2_y y $\hat{S}_{XY}$ estimador insesgado de σ_{xy}). A su vez utilizaremos el estimador reciente obtenido para la razón R. Tenemos:

$$\hat{V}\left(\hat{R}\right) = \frac{1}{\overline{Y}^2 n}\cdot\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right) = \frac{1}{\overline{Y}^2 n(n-1)}\cdot\left[\sum_i^n X_i^2 + \hat{R}^2\sum_i^n Y_i^2 - 2\hat{R}\sum_i^n X_i Y_i\right]$$

ESTIMACIONES DE LOS PARÁMETROS POBLACIONALES BASADAS EN LA RAZÓN Y ERRORES

Podemos utilizar el estimador de la razón para realizar estimaciones de los parámetros poblacionales típicos como sigue:

$$\hat{X}_R = \frac{x}{y}Y = \frac{\bar{x}}{\bar{y}}Y = \hat{R}Y\,,\quad \hat{\bar{X}}_R = \bar{x}_R = \frac{\bar{x}}{\bar{y}}\bar{Y} = \hat{R}\bar{Y}\,,\quad \hat{P}_{RX} = \frac{\hat{P}_X}{\hat{P}_Y}P_Y = \hat{R}P_Y\,,\quad \hat{A}_{RX} = \frac{\hat{A}_X}{\hat{A}_Y}P_Y = \hat{R}A_Y$$

Las varianzas pueden calcularse como sigue:

Muestreo sin reposición

$$V(\hat{X}_R) = V(\hat{R}Y) = Y^2V(\hat{R}) = \; N^2\frac{1-f}{n}\left(S_x^2 + R^2S_y^2 - 2RS_{xy}\right)$$

$$V(\hat{\bar{X}}_R) = V(\hat{R}\bar{Y}) = \bar{Y}^2V(\hat{R}) = \frac{1-f}{n}\left(S_x^2 + R^2S_y^2 - 2RS_{xy}\right)$$

Muestreo con reposición

$$V(\hat{X}_R) = V(\hat{R}Y) = Y^2V(\hat{R}) = \frac{N^2}{n}\left(\sigma_x^2 + R^2\sigma_y^2 - 2R\sigma_{xy}\right)$$

$$V(\hat{\bar{X}}_R) = V(\hat{R}\bar{Y}) = \bar{Y}^2V(\hat{R}) = \frac{1}{n}\left(\sigma_x^2 + R^2\sigma_y^2 - 2R\sigma_{xy}\right)$$

Las estimaciones de las varianzas pueden calcularse como sigue:

Muestreo sin reposición

$$\hat{V}(\hat{X}_R) = N^2\frac{1-f}{n}\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right) = N^2\,\frac{1-f}{n(n-1)}\left[\sum_i^n X_i^2 + \hat{R}^2\sum_i^n Y_i^2 - 2\hat{R}\sum_i^n X_iY_i\right]$$

$$\hat{V}(\hat{\bar{X}}_R) = \frac{1-f}{n}\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right) = \frac{1-f}{n(n-1)}\left[\sum_i^n X_i^2 + \hat{R}^2\sum_i^n Y_i^2 - 2\hat{R}\sum_i^n X_iY_i\right]$$

Muestreo con reposición

$$\hat{V}(\hat{X}_R) = \frac{N^2}{n}\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right) = \frac{N^2}{n(n-1)}\left[\sum_i^n X_i^2 + \hat{R}^2\sum_i^n Y_i^2 - 2\hat{R}\sum_i^n X_iY_i\right]$$

$$\hat{V}(\hat{\bar{X}}_R) = \frac{1}{n}\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right) = \frac{1}{n(n-1)}\left[\sum_i^n X_i^2 + \hat{R}^2\sum_i^n Y_i^2 - 2\hat{R}\sum_i^n X_iY_i\right]$$

ESTIMACIONES POR REGRESIÓN Y ERRORES

Supongamos $(x_i, y_i)\ i = 1,...,N$ pares de valores situados sobre una recta que no pasa por el origen, es decir, $x_i = a + by_i$ con $a \neq 0$. Entonces, para los valores muestrales y poblacionales se cumple, respectivamente $\bar{x} = a + b\bar{y}$ y $\overline{X} = a + b\overline{Y}$ por lo que $\bar{x} - \overline{X} = b\left(\bar{y} - \overline{Y}\right)$, o lo que es lo mismo, $\overline{X} = \bar{x} - b\left(\bar{y} - \overline{Y}\right)$. Se tiene:

- Si $\bar{y} = \overline{Y}$, entonces $\overline{X} = \bar{x}$ y $V(\bar{x}) = 0$
- Si $\bar{y} \neq \overline{Y}$, entonces $\overline{X} \neq \bar{x}$, siendo $b\left(\bar{y} - \overline{Y}\right)$ el ajuste.

Este razonamiento sugiere intentar una ganancia en precisión cuando la relación entre x_i e y_i sea lineal sin pasar por el origen, utilizando el estimador lineal de regresión para la media:

$$\hat{\overline{X}}_{rg} = \bar{x} + b\left(\overline{Y} - \bar{y}\right)$$

Como casos particulares del estimador de regresión se tienen:

1. Si $b = 0$, el estimador de regresión coincide con el estimador directo o de expansión $\left(\hat{\overline{X}}_{rg} = \bar{x}\right)$
2. Si $b = \hat{R} = \dfrac{\bar{x}}{\bar{y}}$, se obtiene el estimador de razón $\left(\hat{\overline{X}}_{rg} = \hat{R}\overline{Y} = \hat{\overline{X}}_R\right)$
3. Si $b = 1$ se obtiene el estimador de la diferencia $\left(\hat{\overline{X}}_{rg} = \bar{x} + \left(\overline{Y} - \bar{y}\right)\right)$

Tenemos:

$$\bar{x}_{rg} = \bar{x} + b_o\left(\overline{Y} - \bar{y}\right) \Rightarrow \begin{cases} b_o = 0 \bar{x}_{rg} = \bar{x}(estimador simple) \\ b_o = \dfrac{\bar{x}}{\bar{y}} \Rightarrow \bar{x}_{rg} = \bar{x} + \dfrac{\bar{x}}{\bar{y}}\left(\overline{Y} - \bar{y}\right) = \bar{x} + \dfrac{\bar{x}}{\bar{y}}\overline{Y} - \dfrac{\bar{x}}{\bar{y}}\bar{y} = \dfrac{\bar{x}}{\bar{y}}\overline{Y} = \hat{\overline{X}}_R(razón) \\ b_o = 1 \bar{x}_{rg} = \left(\bar{x} - \bar{y}\right) + \overline{Y}(estimador\ por\ diferencia) \end{cases}$$

Análogamente, se puede definir el estimador de regresión para el total poblacional como

$$\hat{X}_{rg} = \hat{X} + b\left(Y - \hat{Y}\right)$$

siendo $\hat{X}, \hat{Y}$ los estimadores directos de X, Y respectivamente.

Podemos resumir las estimaciones por regresión como sigue:

$$\bar{x}_{rg} = \bar{x} + b_o\left(\overline{Y} - \bar{y}\right),\ \hat{X}_{rg} = N\bar{x}_{rg},\ \hat{P}_{rg} = \hat{P}_X + b_o\left(P_Y - \hat{P}_Y\right) \text{ y } \hat{A}_{rg} = N\hat{P}_{rg}$$

Sesgo del estimador de regresión

El estimador de regresión es en general sesgado salvo que los puntos (X_i , Y_i) con i = 1,2, ... ,N, donde Y_i representa la variable auxiliar correlacionada con la variable en estudio X_i, estuviesen situados sobre una línea recta que no pasa por el origen de ecuación $X_i = a + b\,Y_i$.

Otro caso de insesgadez del estimador de regresión es cuando $b = b$o = constante.

Varianzas y estimación de varianzas

Las varianzas y sus estimaciones toman los siguientes valores:

Muestreo sin reposición

$$V(\bar{x}_{rg}) = \frac{1-f}{n}\left(S_x^2 + b_o^2 S_y^2 - 2b_o S_{xy}\right), \quad \hat{V}(\bar{x}_{rg}) = \frac{1-f}{n}\left(\hat{S}_x^2 + b_o^2 \hat{S}_y^2 - 2b_o \hat{S}_{xy}\right)$$

$$V(\hat{X}_{rg}) = \frac{N^2(1-f)}{n}\left(S_x^2 + b_o^2 S_y^2 - 2b_o S_{xy}\right), \quad \hat{V}(\hat{X}_{rg}) = \frac{N^2(1-f)}{n}\left(\hat{S}_x^2 + b_o^2 \hat{S}_y^2 - 2b_o \hat{S}_{xy}\right)$$

$$\hat{V}_{\min}(\bar{x}_{rg}) = \frac{1-f}{n}\hat{S}_x^2\left(1-\hat{\rho}^2\right), \quad \hat{V}_{\min}(\hat{X}_{rg}) = \frac{N^2(1-f)}{n}\hat{S}_x^2\left(1-\hat{\rho}^2\right)$$

Muestreo con reposición

$$V(\bar{x}_{rg}) = \frac{1}{n}\left(\sigma_x^2 + b_o^2\sigma_y^2 - 2b_o\sigma_{xy}\right), \quad \hat{V}(\bar{x}_{rg}) = \frac{1}{n}\left(\hat{S}_x^2 + b_o^2\hat{S}_y^2 - 2b_o\hat{S}_{xy}\right), \; \hat{V}_{\min}(\bar{x}_{rg}) = \frac{1}{n}\hat{S}_x^2\left(1-\hat{\rho}^2\right)$$

$$V(\hat{X}_{rg}) = \frac{N^2}{n}\left(\sigma_x^2 + b_o^2\sigma_y^2 - 2b_o\sigma_{xy}\right), \; \hat{V}(\hat{X}_{rg}) = \frac{N^2}{n}\left(\hat{S}_x^2 + b_o^2\hat{S}_y^2 - 2b_o\hat{S}_{xy}\right), \; \hat{V}_{\min}(\hat{X}_{rg}) = \frac{N^2}{n}\hat{S}_x^2\left(1-\hat{\rho}^2\right)$$

Hasta aquí hemos considerado el caso en que b_0 es constante. Sin embargo, cuando se desconoce b_0 o es variable, suelen utilizarse los resultados anteriores, estimando b_0 mediante la expresión:

$$\hat{b}_0 = \hat{\beta} = \frac{\hat{S}_{XY}}{\hat{S}_Y^2} = \frac{\sum_i^n (X_i - \bar{x})(Y_i - \bar{y})}{\sum_i^n (Y_i - \bar{y})^2}$$

Este resutado obtenido es aplicable para muestras grandes.

Comparación con otros tipos de muestreo

Muestreo sin reposición

Para comparar la precisión de la estimación por regresión con la de otros tipos de muestreo utilizamos el estimador de la media y las expresiones de su varianza en los distintos tipos de muestreo. Tenemos:

$$V\left(\hat{\bar{X}}\right) = V(\bar{x}) = \frac{1-f}{n} S_x^2$$

$$V\left(\hat{\bar{X}}_R\right) = \frac{1-f}{n}\left(S_x^2 + R^2 S_y^2 - 2RS_x S_y \cdot \rho_{xy}\right)$$

$$V_{min}\left(\hat{\bar{X}}_{rg}\right) = V_{min}\left(\bar{x}_{rg}\right) = \frac{1-f}{n} S_x^2\left(1 - \rho_{xy}^2\right)$$

Es evidente que $V_{min}\left(\bar{x}_{rg}\right) \leq V(\bar{x})$, ya que $1 - \rho_{xy}^2 \leq 1$, correspondiendo el signo igual al caso $\rho_{xy} = 0$, es decir, al caso de correlación nula entre *X* e *Y*. Por lo tanto, cuando la variable auxiliar y la variable en estudio están incorreladas no se gana en precisión por considerar el método indirecto de estimación por regresión respecto de considerar el muestreo aleatorio simple. En el resto de los casos la estimación indirecta por regresión supera en precisión a la estimación aleatoria simple.

Por otra parte:

$$V_{min}\left(\bar{x}_{rg}\right) < V\left(\bar{x}_R\right) \Leftrightarrow V\left(\bar{x}_R\right) - V_{min}\left(\bar{x}_{rg}\right) \geq 0 \Leftrightarrow$$

$$\frac{1-f}{n}(S_x^2 + R^2 S_y^2 - 2RS_x S_y \rho_{xy}) - \frac{1-f}{n} S_x^2(1 - \rho^2{}_{xy}) \geq 0 \Leftrightarrow$$

$$\frac{1-f}{n}\left(R^2 S_y^2 - 2RS_x S_y \rho_{xy} + S_x^2 \rho_{xy}^2\right) \geq 0 \Leftrightarrow \frac{1-f}{n}\left(RS_y - \rho_{xy} S_x\right)^2 \geq 0$$

La desigualdad es siempre cierta, y se produce la iguladad si:

$$RS_y - \rho_{xy} S_x = 0 \Leftrightarrow R = \rho \frac{S_x}{S_y} = \beta$$

es decir, la igualdad de precisiones en la estimación por razón y por regresión se produce en el caso en que la recta de regresión pase por el origen (si $R = \beta$, la ordenada en el origen de la recta de regresión de *X* sobre *Y*, que en el caso de varianza mínima tiene de ecuación $X = \beta Y + \bar{X} - \beta\bar{Y}$, valdrá $\bar{X} - \beta\bar{Y} = \bar{X} - R\bar{Y} = \bar{X} - \bar{X} = 0$). En cualquier otro caso, la estimación por regresión es más precisa que la estimación por razón.

Muestreo con reposición

Para el caso de muestreo con reposición tenemos:

$$V\left(\hat{\bar{X}}\right) = V(\bar{x}) = \frac{1}{n}\sigma_x^2$$

$$V\left(\hat{\bar{X}}_R\right) = \frac{1}{n}\left(\sigma_x^2 + R^2\sigma_y^2 - 2R\sigma_x\sigma_y \cdot \rho_{xy}\right)$$

$$V_{min}\left(\hat{\bar{X}}_{rg}\right) = V_{min}\left(\bar{x}_{rg}\right) = \frac{1}{n}\sigma_x^2\left(1 - \rho_{xy}^2\right)$$

Es evidente que $V_{min}\left(\bar{x}_{rg}\right) \leq V(\bar{x})$, ya que $1-\rho_{xy}^2 \leq 1$, correspondiendo el signo igual al caso $\rho_{xy}=0$, es decir, al caso de correlación nula entre *X* e *Y*. Por lo tanto, cuando la variable auxiliar y la variable en estudio están incorreladas no se gana en precisión por considerar el método indirecto de estimación por regresión respecto de considerar el muestreo aleatorio simple. En el resto de los casos la estimación indirecta por regresión supera en precisión a la estimación aleatoria simple.

Por otra parte:

$$V_{min}\left(\bar{x}_{rg}\right) < V\left(\bar{x}_R\right) \Leftrightarrow V_{min}\left(\bar{x}_R\right) - V\left(\bar{x}_{rg}\right) \geq 0 \Leftrightarrow$$

$$\frac{1}{n}(\sigma_x^2 + R^2\sigma_y^2 - 2R\sigma_x\sigma_y\rho_{xy}) - \frac{1}{n}\sigma_x^2(1-\rho^2{}_{xy}) \geq 0 \Leftrightarrow$$

$$\frac{1}{n}\left(R^2\sigma_y^2 - 2R\sigma_x\sigma_y\rho_{xy} + \sigma_x^2\rho_{xy}^2\right) \geq 0 \Leftrightarrow \frac{1}{n}\left(R\sigma_y - \rho_{xy}\sigma_x\right)^2 \geq 0$$

La desigualdad es siempre cierta, y se produce la igualdad si:

$$R\sigma_y - \rho_{xy}\sigma_x = 0 \Leftrightarrow R = \rho\frac{\sigma_x}{\sigma_y} = \rho\frac{S_x}{S_y} = \beta$$

es decir, la igualdad de precisiones en la estimación por razón y por regresión se produce en el caso en que la recta de regresión pase por el origen (si $R = \beta$, la ordenada en el origen de la recta de regresión de *X* sobre *Y*, que en el caso de varianza mínima tiene de ecuación $X = \beta Y + \overline{X} - \beta\overline{Y}$, valdrá $\overline{X} - \beta\overline{Y} = \overline{X} - R\overline{Y} = \overline{X} - \overline{X} = 0$). En cualquier otro caso la estimación por regresión es más precisa que la estimación por razón.

ESTIMACIONES POR DIFERENCIA Y ERRORES

Dentro de los denominados métodos indirectos de estimación suele considerarse la estimación por diferencia, que se utiliza en caso de que la recta de regresión que ajusta los puntos (X_i, Y_i) tiene como pendiente la unidad. Por otra parte, ya vimos al estudiar la estimación por regresión que el método de estimación por diferencia era un caso particular suyo (caso en que $b = 1$). Los estimadores de la media y el total basados en el estimador por diferencia $\hat{D} = \bar{x} - \bar{y}$ pueden expresarse como sigue:

$$\hat{\overline{X}} = \bar{x} - \bar{y} + \overline{Y} = \hat{D} + \overline{Y} \qquad \hat{X} = N(\bar{x} - \bar{y}) + Y = \hat{D}_T + Y = N\hat{D} + Y$$

Las varianzas y sus estimaciones para los estimadores de la media y el total basados en la diferencia, coinciden con las varianzas y sus estimaciones de los propios estimadores diferencia. Para muestreo sin reposición tenemos:

$$V(\hat{\overline{X}}) = V(\hat{D} + \overline{Y}) = V(\hat{D}) = \frac{1-f}{n}\left(S_x^2 + S_y^2 - 2S_{xy}\right) \qquad (\overline{Y} \text{ es una constante})$$

$$V(\hat{X}) = V(\hat{D}_T + Y) = V(\hat{D}_T) = N^2 \frac{1-f}{n}\left(S_x^2 + S_y^2 - 2S_{xy}\right) \text{ (Y es una constante)}$$

$$\hat{V}(\hat{\bar{X}}) = \hat{V}\left(\hat{D}\right) = \frac{1-f}{n}\left(\hat{S}_x^2 + \hat{S}_y^2 - 2\hat{S}_{xy}\right), \quad \hat{V}(\hat{X}) = \hat{V}(\hat{D}_T) = N^2 \frac{1-f}{n}\left(\hat{S}_x^2 + \hat{S}_y^2 - 2\hat{S}_{xy}\right)$$

Para muestreo con reposición tenemos:

$$V(\hat{\bar{X}}) = \frac{1}{n}\left(\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}\right) \qquad V(\hat{X}) = N^2 \frac{1}{n}\left(\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}\right)$$

$$\hat{V}(\hat{\bar{X}}) = \hat{V}\left(\hat{D}\right) = \frac{1}{n}\left(\hat{S}_x^2 + \hat{S}_y^2 - 2\hat{S}_{xy}\right) \quad \hat{V}(\hat{X}) = \hat{V}(\hat{D}_T) = N^2 \frac{1}{n}\left(\hat{S}_x^2 + \hat{S}_y^2 - 2\hat{S}_{xy}\right)$$

ESTIMADORES DE RAZÓN EN EL MUESTREO ESTRATIFICADO

Existen dos formas de plantear un estimador de razón para el total X, en el caso de muestreo estratificado. En la primera de ellas se obtiene un estimador de razón para cada el total de cada estrato y se suman todos ellos. El estimador obtenido se denomina *estimador separado de razón*. La principal ventaja de este estimador es que permite que la razón de X a Y varíe de un estrato a otro. Sin embargo, necesitamos conocer el total de la variable auxiliar, Y_h, en cada estrato por separado. En la segunda de ellas se obtiene una única razón con los totales de X e Y estimados mediante muestreo estratificado, es decir, $\frac{\hat{X}_{st}}{\hat{Y}_{st}}$, y se multiplica este cociente por el total de la variable auxiliar Y, que se supone conocido. El estimador obtenido se denomina *estimador combinado de razón*. Para construir este estimador no es preciso conocer el total de la variable auxiliar en cada estrato; basta conocer el total de la población. Ésta es una ventaja con respecto al estimador separado de razón. Sin embargo, el estimador combinado supone, implícitamente, que la razón permanece constante de un estrato a otro.

Estimador de razón simple o separado (para el total poblacional)

Se define el estimador separado de razón para el total poblacional X en un muestreo estratificado como:

$$\hat{X}_{RS} = \sum_{h=1}^{L} \hat{X}_{Rh} = \sum_{h=1}^{L} \hat{R}_h Y_h = \sum_{h=1}^{L} \frac{\bar{x}_h}{\bar{y}_h} Y_h$$

Se observa que es la suma de los estimadores de razón para el total en los diferentes estratos. En general este estimador es sesgado, por serlo $\hat{R}_h \ \forall h = 1,...,L$.

Sesgo del estimador de razón simple o separado y su estimación

Muestreo sin reposición

$$B(\hat{X}_{RS}) = E(\hat{X}_{RS}) - X = E(\sum_h^L \hat{R}_h Y_h) - \sum_h^L X_h = \sum_h^L E(\hat{R}_h) Y_h - \sum_h^L \frac{X_h}{Y_h} Y_h =$$

$$\sum_h^L E(\hat{R}_h) Y_h - \sum_h^L R_h Y_h = \sum_h^L (\underbrace{E(\hat{R}_h) - R_h}_{B(\hat{R}_h)}) Y_h = \sum_h^L B(\hat{R}_h) Y_h$$

Se observa que el sesgo total es la suma de los sesgos en cada estrato ponderados por los Y_h. *Para muestreo sin reposición la expresión del sesgo será*:

$$B(\hat{X}_{RS}) = \sum_h^L Y_h B(\hat{R}_h) = \sum_h^L Y_h \frac{(1-f_h)}{n_h \underbrace{\overline{Y}_h^2}_{Y_h^2/N_h^2}} \left(R_h S_{Yh}^2 - S_{XYh}\right) = \sum_h^L \frac{N_h^2(1-f_h)}{n_h Y_h} \left(R_h S_{Yh}^2 - S_{XYh}\right)$$

que *puede estimarse como*: $\hat{B}(\hat{X}_{RS}) = \sum_h^L \frac{N_h^2(1-f_h)}{n_h Y_h} \left(\hat{R}_h \hat{S}_{Yh}^2 - \hat{S}_{XYh}\right)$

Muestreo con reposición

Para muestreo con reposición la expresión del sesgo será:

$$B(\hat{X}_{RS}) = \sum_h^L Y_h B(\hat{R}_h) = \sum_h^L Y_h \frac{1}{n_h \underbrace{\overline{Y}_h^2}_{Y_h^2/N_h^2}} \left(R_h \sigma_{Yh}^2 - \sigma_{XYh}\right) = \sum_h^L \frac{N_h^2}{n_h Y_h} \left(R_h \sigma_{Yh}^2 - \sigma_{XYh}\right)$$

que *puede estimarse como*: $\hat{B}(\hat{X}_{RS}) = \sum_h^L \frac{N_h^2}{n_h Y_h} \left(\hat{R}_h \hat{S}_{Yh}^2 - \hat{S}_{XYh}\right)$

Varianza del estimador de razón simple o separado y su estimación

Muestreo sin reposición

El valor de la *varianza de este estimador para muestreo sin reposición* será:

$$V(\hat{X}_{RS}) = \sum_h^L V(\hat{R}_h \cdot Y_h) = \sum_h^L Y_h^2 \cdot V(\hat{R}_h) = \sum_h^L \underbrace{Y_h^2}_{\downarrow N_h^2 \overline{Y}_h^2} \cdot \frac{1-f_h}{\overline{Y}_h^2 n_h} \left(S_{xh}^2 + R_h^2 S_{yh}^2 - 2R_h S_{xyh}\right) =$$

$$\sum_h^L \frac{N_h^2(1-f_h)}{n_h} \left(S_{xh}^2 + R_h^2 S_{yh}^2 - 2R_h S_{xyh}\right) = \sum_h^L \frac{N_h^2(1-f_h)}{n_h(N_h-1)} \left(\sum_i^{N_h} X_{hi}^2 + R_h^2 \sum_i^{N_h} Y_{hi}^2 - 2R_h \sum_i^{N_h} X_{hi} Y_{hi}\right)$$

La *estimación de la varianza para muestreo sin reposición* será:

$$\hat{V}(\hat{X}_{RS}) = \sum_h^L \frac{N_h^2(1-f_h)}{n_h} \left(\hat{S}_{xh}^2 + \hat{R}_h^2 \hat{S}_{yh}^2 - 2\hat{R}_h \hat{S}_{xyh}\right) =$$

$$= \sum_h^L \frac{N_h^2(1-f_h)}{n_h(n_h-1)} \left(\sum_i^{n_h} X_{hi}^2 + \hat{R}_h^2 \sum_i^{n_h} Y_{hi}^2 - 2\hat{R}_h \sum_i^{n_h} X_{hi} Y_{hi}\right)$$

Muestreo con reposición

El valor de la *varianza del estimador separado del total para muestreo con reposición* será:

$$V(\hat{X}_{RS}) = \sum_{h}^{L} V(\hat{R}_h \cdot Y_h) = \sum_{h}^{L} Y_h^2 \cdot V(\hat{R}_h) = \sum_{h}^{L} \underset{\downarrow}{Y_h^2} \cdot \frac{1}{\overline{Y}_h^2 n_h}\left(\sigma_{xh}^2 + R_h^2\sigma_{yh}^2 - 2R_h\sigma_{xyh}\right) =$$
$$N_h^2\overline{Y}_h^2$$

$$\sum_{h}^{L} \frac{N_h^2}{n_h}\left(\sigma_{xh}^2 + R_h^2\sigma_{yh}^2 - 2R_h\sigma_{xyh}\right) = \sum_{h}^{L} \frac{N_h^2}{n_h(N_h-1)}\left(\sum_{i}^{N_h} X_{hi}^2 + R_h^2\sum_{i}^{N_h} Y_{hi}^2 - 2R_h\sum_{i}^{N_h} X_{hi}Y_{hi}\right)$$

La *estimación de la varianza para muestreo con reposición* será:

$$\hat{V}(\hat{X}_{RS}) = \sum_{h}^{L} \frac{N_h^2}{n_h}\left(\hat{S}_{xh}^2 + \hat{R}_h^2\hat{S}_{yh}^2 - 2\hat{R}_h\hat{S}_{xyh}\right) = \sum_{h}^{L} \frac{N_h^2}{n_h(n_h-1)}\left(\sum_{i}^{n_h} X_{hi}^2 + \hat{R}_h^2\sum_{i}^{n_h} Y_{hi}^2 - 2\hat{R}_h\sum_{i}^{n_h} X_{hi}Y_{hi}\right)$$

Estimador de razón simple o separado (para la media poblacional)

Se consideran estimaciones para la media basadas en la razón en cada estrato definidas como $\hat{\overline{X}}_{Rh} = \frac{\overline{x}_h}{\overline{y}_h} \cdot \overline{Y}_h = \hat{R}_h \cdot \overline{Y}_h$. Como en muestreo estratificado la estimación del total se forma sumando las estimaciones de las medias en cada estrato ponderadas por los $W_h = N_h/N$ ($\hat{\overline{X}}st = \sum_{h=1}^{L} W_h \hat{\overline{X}}_h$), podemos definir el estimador simple o separado de la media como:

$$\hat{\overline{X}}_{RS} = \sum_{h}^{L} W_h \hat{\overline{X}}_{Rh} = \sum_{h}^{L} W_h \hat{R}_h \cdot \overline{Y}_h$$

Este estimador para la media puede expresarse como:

$$\hat{\overline{X}}_{RS} = \sum_{h}^{L} W_h \hat{\overline{X}}_{Rh} = \sum_{h}^{L} W_h \hat{R}_h \cdot \overline{Y}_h = \sum_{h}^{L} \frac{N_h}{N}\hat{R}_h \cdot \frac{Y_h}{N_h} = \frac{1}{N}\sum_{h}^{L} \hat{R}_h Y_h = \frac{\hat{X}_{RS}}{N}$$

Luego todas las fórmulas para el estimador de la media pueden obtenerse a partir de las fórmulas correspondientes ya vistas para el estimador del total.

Muestreo sin reposición

El valor de la *varianza de este estimador para muestreo sin reposición* será:

$$V(\hat{\overline{X}}_{RS}) = \frac{1}{N^2}V(\hat{X}_{RS}) = \sum_{h}^{L} \underbrace{\left(\frac{N_h^2}{N^2}\right)}_{W_h^2} \frac{(1-f_h)}{n_h}\left(S_{xh}^2 + R_h^2 S_{yh}^2 - 2R_h S_{xyh}\right) =$$

$$\sum_{h}^{L} \frac{W_h^2(1-f_h)}{n_h(N_h-1)}\left(\sum_{i}^{N_h} X_{hi}^2 + R_h^2\sum_{i}^{N_h} Y_{hi}^2 - 2R_h\sum_{i}^{N_h} X_{hi}Y_{hi}\right)$$

La *estimación de la varianza para muestreo sin reposición* será:

$$\hat{V}(\hat{\overline{X}}_{RS}) = \sum_{h}^{L} \frac{W_h^2(1-f_h)}{n_h}\left(\hat{S}_{xh}^2 + \hat{R}_h^2\hat{S}_{yh}^2 - 2\hat{R}_h\hat{S}_{xyh}\right) =$$

$$= \sum_{h}^{L} \frac{W_h^2(1-f_h)}{n_h(n_h-1)}\left(\sum_{i}^{n_h} X_{hi}^2 + \hat{R}_h^2\sum_{i}^{n_h} Y_{hi}^2 - 2\hat{R}_h\sum_{i}^{n_h} X_{hi}Y_{hi}\right)$$

El valor del *sesgo del estimador simple o separado* es el siguiente:

$$B(\hat{\bar{X}}_{RS}) = E(\hat{\bar{X}}_{RS}) - \bar{X} = E(\frac{\hat{X}_{RS}}{N}) - \frac{X}{N} = \frac{1}{N}(E(\hat{X}_{RS}) - X) = \frac{1}{N}B(\hat{X}_{RS}) = \sum_{h}^{L} B(\hat{R}_h)\frac{Y_h}{N}$$

Se observa que *el sesgo total es la suma de los sesgos en cada estrato* ponderados por los Y_h/N. *Para muestreo sin reposición la expresión del sesgo será*:

$$B(\hat{\bar{X}}_{RS}) = \frac{1}{N}B(\hat{X}_{RS}) = \sum_{h}^{L}\frac{N_h^2(1-f_h)}{Nn_hY_h}\left(R_hS_{Yh}^2 - S_{XYh}\right) = \sum_{h}^{L}\frac{W_h(1-f_h)}{n_h\bar{Y}_h}\left(R_hS_{Yh}^2 - S_{XYh}\right)$$

que *puede estimarse como*: $\hat{B}(\hat{\bar{X}}_{RS}) = \sum_{h}^{L}\frac{W_h(1-f_h)}{n_h\bar{Y}_h}\left(\hat{R}_h\hat{S}_{Yh}^2 - \hat{S}_{XYh}\right)$

Muestreo con reposición

El valor de la *varianza del estimador separado de la media para muestreo con reposición* será:

$$V(\hat{\bar{X}}_{RS}) = \frac{1}{N^2}V(\hat{X}_{RS}) = \sum_{h}^{L}\underbrace{\left(\frac{N_h^2}{N^2}\right)}_{W_h^2}\frac{1}{n_h}\left(\sigma_{xh}^2 + R_h^2\sigma_{yh}^2 - 2R_h\sigma_{xyh}\right) =$$

$$\sum_{h}^{L}\frac{W_h^2}{n_hN_h}\left(\sum_{i}^{N_h}X_{hi}^2 + R_h^2\sum_{i}^{N_h}Y_{hi}^2 - 2R_h\sum_{i}^{N_h}X_{hi}Y_{hi}\right)$$

La *estimación de la varianza para muestreo con reposición* será:

$$\hat{V}(\hat{\bar{X}}_{RS}) = \sum_{h}^{L}\frac{W_h^2}{n_h}\left(\hat{S}_{xh}^2 + \hat{R}_h^2\hat{S}_{yh}^2 - 2\hat{R}_h\hat{S}_{xyh}\right) = \sum_{h}^{L}\frac{W_h^2}{n_h(n_h-1)}\left(\sum_{i}^{n_h}X_{hi}^2 + \hat{R}_h^2\sum_{i}^{n_h}Y_{hi}^2 - 2\hat{R}_h\sum_{i}^{n_h}X_{hi}Y_{hi}\right)$$

Para muestreo con reposición la expresión del sesgo será:

$$B(\hat{\bar{X}}_{RS}) = \frac{1}{N}B(\hat{X}_{RS}) = \sum_{h}^{L}\frac{N_h^2}{Nn_hY_h}\left(R_h\sigma_{Yh}^2 - \sigma_{XYh}\right) = \sum_{h}^{L}\frac{W_h}{n_h\bar{Y}_h}\left(R_h\sigma_{Yh}^2 - \sigma_{XYh}\right)$$

que *puede estimarse como*: $\hat{B}(\hat{X}_{RS}) = \sum_{h}^{L}\frac{W_h}{n_h\bar{Y}_h}\left(\hat{R}_h\hat{S}_{Yh}^2 - \hat{S}_{XYh}\right)$

El método de estimación estratificada por razón simple o separada presenta como *principal ventaja* la obtención de estimaciones separadas por estratos, lo que permite ofrecer información de la población al subnivel de estratos. El *principal inconveniente* de este método es la acumulación de los sesgos de las estimaciones en los estratos para el cálculo del sesgo total. *En la práctica suele utilizarse este método cuando* los estratos son de tamaño elevado (habrá pocos estratos en la población, lo que implica pocos sumandos en la acumulación de sesgos). También suele utilizarse cuando los R_h tienden a ser muy distintos.

Estimador de razón combinado (para el total poblacional)

Se considera inicialmente la razón de los estimadores estratificados $\hat{R}_C = \dfrac{\bar{x}_{st}}{\bar{y}_{st}} = \dfrac{\hat{\bar{X}}_{st}}{\hat{\bar{Y}}_{st}}$, y se forma el estimador del total $\hat{X}_{RC} = \hat{R}_C \cdot Y$ (ya que el estimador del total basado en la razón es $\hat{X} = \hat{R} \cdot Y$).

Muestreo sin reposición

El valor de la *varianza de este estimador para muestreo sin reposición* será:

$$V(\hat{X}_{RC}) = V(\hat{R}_C \cdot Y) = Y^2 \cdot V(\hat{R}_C) = \underbrace{Y^2}_{N^2\bar{Y}^2} \cdot \frac{1}{\bar{Y}^2} \Big(\underbrace{V(\bar{x}_{st})}_{\sum_h^L W_h^2(1-f_h)\frac{S_{Xh}^2}{n_h}} + R^2 \underbrace{V(\bar{y}_{st})}_{\sum_h^L W_h^2(1-f_h)\frac{S_{Yh}^2}{n_h}} - 2R \underbrace{Cov(\bar{x}_{st}, \bar{y}_{st})}_{\sum_h^L W_h^2(1-f_h)\frac{S_{XYh}}{n_h}} \Big)$$

$$N^2 \sum_h^L \frac{W_h^2(1-f_h)}{n_h} \left(S_{xh}^2 + R^2 S_{yh}^2 - 2R\, S_{xyh} \right) = N^2 \sum_h^L \frac{W_h^2(1-f_h)}{n_h(N_h-1)} \left(\sum_i^{N_h} X_{hi}^2 + R^2 \sum_i^{N_h} Y_{hi}^2 - 2R \sum_i^{N_h} X_{hi} Y_{hi} \right)$$

En el cálculo de esta varianza se ha aplicado la fórmula general de la varianza del estimador de la razón ya estudiada anteriormente.

La *estimación de la varianza para muestreo sin reposición* será:

$$\hat{V}(\hat{X}_{RC}) = N^2 \sum_h^L \frac{W_h^2(1-f_h)}{n_h} \left(\hat{S}_{xh}^2 + \hat{R}^2 \hat{S}_{yh}^2 - 2\hat{R}\, \hat{S}_{xyh} \right) = N^2 \sum_h^L \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \left(\sum_i^{n_h} X_{hi}^2 + R^2 \sum_i^{n_h} Y_{hi}^2 - 2R \sum_i^{n_h} X_{hi} Y_{hi} \right)$$

El valor del *sesgo del estimador combinado para el total* es el siguiente:

$$B(\hat{X}_{RC}) = E(\hat{X}_{RC}) - X = E(\hat{R}_C Y) - \frac{X}{Y} Y = E(\hat{R}_C) Y - RY = \left(E(\hat{R}_C) - R \right) Y = B(\hat{R}_C) Y$$

Se observa que para el sesgo total no se acumulan los sesgos en cada estrato. *Para muestreo sin reposición la expresión del sesgo será*:

$$B(\hat{X}_{RC}) = B(\hat{R}_C) Y = \frac{R \overbrace{V(\bar{y}_{st})}^{\sum_h^L W_h^2(1-f_h)\frac{S_{Yh}^2}{n_h}} - \overbrace{Cov(\bar{x}_{st}, \bar{y}_{st})}^{\sum_h^L W_h^2(1-f_h)\frac{S_{XYh}}{n_h}}}{\underbrace{\bar{Y}^2}_{Y^2/N^2}} \cdot Y = N^2 \sum_h^L \frac{W_h^2(1-f_h)}{n_h Y} (RS_{Yh}^2 - S_{XYh})$$

que *puede estimarse como*: $\hat{B}(\hat{X}_{RC}) = N^2 \sum_h^L \dfrac{W_h^2(1-f_h)}{n_h Y} \left(\hat{R}\hat{S}_{Yh}^2 - \hat{S}_{XYh} \right)$

Muestreo con reposición

El valor de la *varianza del estimador combinado del total para muestreo con reposición* será:

$$V(\hat{X}_{RC}) = V(\hat{R}_C \cdot Y) = Y^2 \cdot V(\hat{R}_C) = \underbrace{Y^2}_{N^2\overline{Y}^2} \cdot \frac{1}{\overline{Y}^2}\left(\underbrace{V(\overline{x}_{st})}_{\sum_h^L W_h^2 \frac{\sigma_{Xh}^2}{n_h}} + R^2 \underbrace{V(\overline{y}_{st})}_{\sum_h^L W_h^2 \frac{\sigma_{Yh}^2}{n_h}} - 2R\underbrace{Cov(\overline{x}_{st}, \overline{y}_{st})}_{\sum_h^L W_h^2 \frac{\sigma_{XYh}}{n_h}}\right)$$

$$N^2 \sum_h^L \frac{W_h^2}{n_h}\left(\sigma_{xh}^2 + R^2\sigma_{yh}^2 - 2R\ \sigma_{xyh}\right) = N^2 \sum_h^L \frac{W_h^2}{n_h N_h}\left(\sum_i^{N_h} X_{hi}^2 + R^2 \sum_i^{N_h} Y_{hi}^2 - 2R\sum_i^{N_h} X_{hi}Y_{hi}\right)$$

La *estimación de la varianza para muestreo con reposición* será:

$$\hat{V}(\hat{X}_{RC}) = N^2 \sum_h^L \frac{W_h^2}{n_h}\left(\hat{S}_{xh}^2 + \hat{R}^2\hat{S}_{yh}^2 - 2\hat{R}\hat{S}_{xyh}\right) = N^2 \sum_h^L \frac{W_h^2}{n_h(n_h-1)}\left(\sum_i^{n_h} X_{hi}^2 + \hat{R}^2 \sum_i^{n_h} Y_{hi}^2 - 2\hat{R}\sum_i^{n_h} X_{hi}Y_{hi}\right)$$

Para muestreo con reposición la expresión del sesgo será:

$$B(\hat{X}_{RC}) = B(\hat{R}_C)Y = \frac{R\overbrace{V(\overline{y}_{st})}^{\sum_h^L W_h^2 \frac{\sigma_{Yh}^2}{n_h}} - \overbrace{Cov(\overline{x}_{st}, \overline{y}_{st})}^{\sum_h^L W_h^2 \frac{\sigma_{XYh}}{n_h}}}{\underbrace{\overline{Y}^2}_{Y^2/N^2}} \cdot Y = N^2 \sum_h^L \frac{W_h^2}{n_h Y}(R\sigma_{Yh}^2 - \sigma_{XYh})$$

que *puede estimarse como*: $\hat{B}(\hat{X}_{RC}) = N^2 \sum_h^L \frac{W_h^2}{n_h Y}\left(\hat{R}\hat{S}_{Yh}^2 - \hat{S}_{XYh}\right)$

Estimador de razón combinado (para la media poblacional)

Se considera inicialmente la razón de los estimadores estratificados $\hat{R}_C = \frac{\overline{x}_{st}}{\overline{y}_{st}} = \frac{\hat{\overline{X}}_{st}}{\hat{\overline{Y}}_{st}}$, y se forma el estimador de la media $\hat{\overline{X}}_{RC} = \hat{R}_C \cdot \overline{Y}$ (ya que el estimador del total basado en la razón es $\hat{\overline{X}} = \hat{R} \cdot \overline{Y}$).

Muestreo sin reposición

El valor de la *varianza de este estimador para muestreo sin reposición* será:

$$V(\hat{\overline{X}}_{RC}) = V(\hat{R}_C \cdot \overline{Y}) = \overline{Y}^2 \cdot V(\hat{R}_C) = \overline{Y}^2 \cdot \frac{1}{\overline{Y}^2}\left(\underbrace{V(\overline{x}_{st})}_{\sum_h^L W_h^2(1-f_h)\frac{S_{Xh}^2}{n_h}} + R^2 \underbrace{V(\overline{y}_{st})}_{\sum_h^L W_h^2(1-f_h)\frac{S_{Yh}^2}{n_h}} - 2R\underbrace{Cov(\overline{x}_{st}, \overline{y}_{st})}_{\sum_h^L W_h^2(1-f_h)\frac{S_{XYh}}{n_h}}\right)$$

$$\sum_h^L \frac{W_h^2(1-f_h)}{n_h}\left(S_{xh}^2 + R^2 S_{yh}^2 - 2R\ S_{xyh}\right) = \sum_h^L \frac{W_h^2(1-f_h)}{n_h(N_h-1)}\left(\sum_i^{N_h} X_{hi}^2 + R^2 \sum_i^{N_h} Y_{hi}^2 - 2R\sum_i^{N_h} X_{hi}Y_{hi}\right)$$

En el cálculo de esta varianza se ha aplicado la fórmula general de la varianza del estimador de la razón ya estudiada anteriormente.

La *estimación de la varianza para muestreo sin reposición* será:

$$\hat{V}(\hat{\overline{X}}_{RC}) = \sum_h^L \frac{W_h^2(1-f_h)}{n_h}\left(\hat{S}_{xh}^2 + \hat{R}^2\hat{S}_{yh}^2 - 2\hat{R}\,\hat{S}_{xyh}\right) = \sum_h^L \frac{W_h^2(1-f_h)}{n_h(n_h-1)}\left(\sum_i^{n_h} X_{hi}^2 + R^2 \sum_i^{n_h} Y_{hi}^2 - 2R\sum_i^{n_h} X_{hi}Y_{hi}\right)$$

El valor del *sesgo del estimador combinado para la media* es el siguiente:

$$B(\hat{\bar{X}}_{RC}) = E(\hat{\bar{X}}_{RC}) - \bar{X} = E(\hat{R}_C\bar{Y}) - \frac{\bar{X}}{\bar{Y}}\bar{Y} = E(\hat{R}_C)\bar{Y} - R\bar{Y} = \left(E(\hat{R}_C) - R\right)\bar{Y} = B(\hat{R}_C)\bar{Y}$$

Se observa que para el sesgo total no se acumulan los sesgos en cada estrato. *Para muestreo sin reposición la expresión del sesgo será*:

$$B(\hat{\bar{X}}_{RC}) = B(\hat{R}_C)\bar{Y} = \frac{R\overbrace{V(\bar{y}_{st})}^{\sum_h^L W_h^2(1-f_h)\frac{S_{Yh}^2}{n_h}} - \overbrace{Cov(\bar{x}_{st},\bar{y}_{st})}^{\sum_h^L W_h^2(1-f_h)\frac{S_{XYh}}{n_h}}}{\bar{Y}^2}\cdot\bar{Y} = \sum_h^L \frac{W_h^2(1-f_h)}{n_h\bar{Y}}(RS_{Yh}^2 - S_{XYh})$$

que *puede estimarse como*: $\hat{B}(\hat{\bar{X}}_{RC}) = \sum_h^L \frac{W_h^2(1-f_h)}{n_h Y}\left(\hat{R}\hat{S}_{Yh}^2 - \hat{S}_{XYh}\right)$

Muestreo con reposición

El valor de la *varianza del estimador combinado de la media para muestreo con reposición* será:

$$V(\hat{\bar{X}}_{RC}) = V(\hat{R}_C\cdot\bar{Y}) = \bar{Y}^2\cdot V(\hat{R}_C) = \bar{Y}^2\cdot\frac{1}{\bar{Y}^2}\left(\underbrace{V(\bar{x}_{st})}_{\sum_h^L W_h^2\frac{\sigma_{Xh}^2}{n_h}} + R^2\underbrace{V(\bar{y}_{st})}_{\sum_h^L W_h^2\frac{\sigma_{Yh}^2}{n_h}} - 2R\underbrace{Cov(\bar{x}_{st},\bar{y}_{st})}_{\sum_h^L W_h^2\frac{\sigma_{XYh}}{n_h}}\right)$$

$$\sum_h^L \frac{W_h^2}{n_h}\left(\sigma_{xh}^2 + R^2\sigma_{yh}^2 - 2R\,\sigma_{xyh}\right) = \sum_h^L \frac{W_h^2}{n_h N_h}\left(\sum_i^{N_h} X_{hi}^2 + R^2\sum_i^{N_h} Y_{hi}^2 - 2R\sum_i^{N_h} X_{hi}Y_{hi}\right)$$

La *estimación de la varianza para muestreo con reposición* será:

$$\hat{V}(\hat{\bar{X}}_{RC}) = \sum_h^L \frac{W_h^2}{n_h}\left(\hat{S}_{xh}^2 + \hat{R}^2\hat{S}_{yh}^2 - 2\hat{R}\hat{S}_{xyh}\right) = \sum_h^L \frac{W_h^2}{n_h(n_h-1)}\left(\sum_i^{n_h} X_{hi}^2 + \hat{R}^2\sum_i^{n_h} Y_{hi}^2 - 2\hat{R}\sum_i^{n_h} X_{hi}Y_{hi}\right)$$

Para muestreo con reposición la expresión del sesgo será:

$$B(\hat{\bar{X}}_{RC}) = B(\hat{R}_C)\bar{Y} = \frac{R\overbrace{V(\bar{y}_{st})}^{\sum_h^L W_h^2\frac{\sigma_{Yh}^2}{n_h}} - \overbrace{Cov(\bar{x}_{st},\bar{y}_{st})}^{\sum_h^L W_h^2\frac{\sigma_{XYh}}{n_h}}}{\bar{Y}^2}\cdot\bar{Y} = \sum_h^L \frac{W_h^2}{n_h\bar{Y}}(R\sigma_{Yh}^2 - \sigma_{XYh})$$

que *puede estimarse como*: $\hat{B}(\hat{\bar{X}}_{RC}) = \sum_h^L \frac{W_h^2}{n_h\bar{Y}}\left(\hat{R}\hat{S}_{Yh}^2 - \hat{S}_{XYh}\right)$

El método de estimación estratificada por razón combinada presenta como *principal ventaja* la no acumulación de los sesgos de las estimaciones en los estratos para el cálculo del sesgo total, lo que reduce el sesgo del estimador final respecto de la estimación separada. El *principal inconveniente* de este método es la imposibilidad de obtención de estimaciones separadas por estratos, lo que no permite disponer de información de la población al subnivel de estratos. *En la práctica suele utilizarse este método cuando* los estratos son de tamaño pequeño (habrá muchos estratos en la población, lo que implica demasiado sesgo por estimación separada). En general suele utilizarse siempre que la estimación separada presenta demasiado sesgo. También suele utilizarse cuando los R_h tienden a ser constantes.

ESTIMADORES DE REGRESIÓN EN EL MUESTREO ESTRATIFICADO

También distinguiremos aquí entre el estimador simple o separado obtenido a partir de estimaciones de regresión en cada estrato, cuya expresión será $\bar{x}_{rgst} = \sum_h^L W_h \bar{x}_{rgh}$, y el estimador combinado, obtenido directamente a partir de las medias estratificadas, que vale $\bar{x}_{rgc} = \bar{x}_{st} + b\left(\bar{Y} - \bar{y}_{st}\right)$.

Ambos estimadores son insesgados para un valor b_o prefijado de b, ya que:

$$E(\bar{x}_{rgst}) = \sum_h^L W_h E(\bar{x}_{rgh}) = \sum_h^L W_h \bar{X}_h = \bar{X}$$
$$E(\bar{x}_{rgc}) = E(\bar{x}_{st}) + b\left(\bar{Y} - E(\bar{y}_{st})\right) = \bar{X} + b(\bar{Y} - \bar{Y}) = \bar{X}$$

Como en el caso de los estimadores de la razón, el estimador combinado suele ser más apropiado que el simple cuando el sesgo de $\bar{x}_{rgh}$ es aproximadamente constante en los diversos estratos y esperamos regresiones lineales en ellos.

Estimador simple o separado

Muestreo sin reposición

En el supuesto $b = b_o$ la varianza del estimador simple para la media es:

$$V\left(\bar{x}_{rgst}\right) = \sum_h^L W_h^2 V\left(\bar{x}_{rgh}\right) = \sum_h^L W_h^2 \frac{1-f_h}{n_h}(S_{Xh}^2 + b_o S_{Xh}^2 - 2b_o S_{XYh})$$

que será mínima cuando lo sean las $V\left(\bar{x}_{rgh}\right)$, es decir, cuando $b_o = \beta_h = \frac{S_{XYh}}{S_{Yh}^2}$

La varianza mínima será entonces:

$$V\left(\bar{x}_{rgst}\right) = \sum_h^L W_h^2 V\left(\bar{x}_{rgh}\right) = \sum_h^L W_h^2 \frac{1-f_h}{n_h}(S_{Xh}^2 + \beta_h S_{Xh}^2 - 2\beta_h S_{XYh}) = \sum_h^L W_h^2 \frac{1-f_h}{n_h} S_{xh}^2 (1-\rho^2{}_{xyh})$$

que puede estimarse mediante:

$$\hat{V}\left(\bar{x}_{rgst}\right) = \sum_h^L W_h^2 \frac{1-f_h}{n_h}(\hat{S}_{Xh}^2 + \hat{\beta}_h \hat{S}_{Xh}^2 - 2\hat{\beta}_h \hat{S}_{XYh}) = \sum_h^L W_h^2 \frac{1-f_h}{n_h} \hat{S}_{xh}^2 (1-\hat{\rho}^2{}_{xyh})$$

Para la estimación separada del total $\hat{X}_{rgst} = \sum_h^L N_h \bar{x}_{rgh}$ se tiene:

$$V\left(\hat{X}_{rgst}\right) = \sum_h^L N_h^2 V\left(\bar{x}_{rgh}\right) = \sum_h^L N_h^2 \frac{1-f_h}{n_h}(S_{Xh}^2 + \beta_h S_{Xh}^2 - 2\beta_h S_{XYh}) = \sum_h^L N_h^2 \frac{1-f_h}{n_h} S_{xh}^2 (1-\rho^2{}_{xyh})$$

que puede estimarse mediante:

$$\hat{V}\left(\hat{X}_{rgst}\right) = \sum_{h}^{L} N_h^2 \frac{1-f_h}{n_h}(\hat{S}_{Xh}^2 + \hat{\beta}_h \hat{S}_{Xh}^2 - 2\hat{\beta}_h \hat{S}_{XYh}) = \sum_{h}^{L} N_h^2 \frac{1-f_h}{n_h} \hat{S}_{xh}^2 (1 - \hat{\rho}^2{}_{xyh})$$

Muestreo con reposición

En el supuesto $b = b_o$ la varianza del estimador simple es:

$$V\left(\bar{x}_{rgst}\right) = \sum_{h}^{L} W_h^2 V\left(\bar{x}_{rgh}\right) = \sum_{h}^{L} W_h^2 \frac{1}{n_h}(\sigma_{Xh}^2 + b_o \sigma_{Xh}^2 - 2b_o \sigma_{XYh})$$

que será mínima cuando lo sean las $V\left(\bar{x}_{rgh}\right)$, es decir, cuando $b_o = \beta_h = \dfrac{S_{XYh}}{S_{Yh}^2} = \dfrac{\sigma_{XYh}}{\sigma_{Yh}^2}$

La varianza mínima será entonces:

$$V_{min}\left(\bar{x}_{rgst}\right) = \sum_{h}^{L} W_h^2 V\left(\bar{x}_{rgh}\right) = \sum_{h}^{L} W_h^2 \frac{1}{n_h}(\sigma_{Xh}^2 + \beta_h \sigma_{Xh}^2 - 2\beta_h \sigma_{XYh}) = \sum_{h}^{L} W_h^2 \frac{1}{n_h} \sigma_{xh}^2 (1 - \rho^2{}_{xyh})$$

que puede estimarse mediante:

$$\hat{V}_{min}\left(\bar{x}_{rgst}\right) = \sum_{h}^{L} W_h^2 \frac{1}{n_h}(\hat{S}_{Xh}^2 + \hat{\beta}_h \hat{S}_{Xh}^2 - 2\hat{\beta}_h \hat{S}_{XYh}) = \sum_{h}^{L} W_h^2 \frac{1}{n_h} \hat{S}_{xh}^2 (1 - \hat{\rho}^2{}_{xyh})$$

Para el estimador del total se tendría:

$$V_{min}\left(\hat{X}_{rgst}\right) = \sum_{h}^{L} N_h^2 V\left(\bar{x}_{rgh}\right) = \sum_{h}^{L} N_h^2 \frac{1}{n_h}(\sigma_{Xh}^2 + \beta_h \sigma_{Xh}^2 - 2\beta_h \sigma_{XYh}) = \sum_{h}^{L} N_h^2 \frac{1}{n_h} \sigma_{xh}^2 (1 - \rho^2{}_{xyh})$$

$$\hat{V}_{min}\left(\hat{X}_{rgst}\right) = \sum_{h}^{L} N_h^2 \frac{1}{n_h}(\hat{S}_{Xh}^2 + \hat{\beta}_h \hat{S}_{Xh}^2 - 2\hat{\beta}_h \hat{S}_{XYh}) = \sum_{h}^{L} N_h^2 \frac{1}{n_h} \hat{S}_{xh}^2 (1 - \hat{\rho}^2{}_{xyh})$$

Estimador combinado

Muestreo sin reposición

El estimador combinado para la media se forma como:

$$\bar{x}_{rgc} = \bar{x}_{st} + b_o\left(\bar{Y} - \bar{y}_{st}\right) \quad con \quad \bar{x}_{st} = \sum_{h}^{L} W_h \bar{x}_h \quad \bar{y}_{st} = \sum_{h}^{L} W_h \bar{y}_h$$

Su varianza puede expresarse de la siguiente forma:

$$V\left(\bar{x}_{rgc}\right) = V\left(\bar{x}_{st}\right) + b_o^2 V\left(\bar{Y} - \bar{y}_{st}\right) - 2b_o \operatorname{cov}\left(\bar{x}_{st}, \bar{Y} - \bar{y}_{st}\right) =$$

$$V\left(\bar{x}_{st}\right) + b_o^2 V\left(\bar{y}_{st}\right) - 2b_o \operatorname{cov}\left(\bar{x}_{st}, \bar{y}_{st}\right) = \sum_{h}^{L} \frac{W_h^2\left(1-f_h\right)}{n_h} \cdot \left(S_{xh}^2 + b_o^2 S_{yh}^2 - 2b_o S_{xyh}\right)$$

Para hallar el valor de b_o que minimiza esta expresión, igualamos a cero su derivada respecto de b_o y tenemos:

$$2b_0\sum_h^L \frac{W_h^2\left(1-f_h\right)}{n_h}\cdot S_{yh}^2 - 2\sum_h^L \frac{W_h^2\left(1-f_h\right)}{n_h}\cdot S_{xyh} = 0 \Rightarrow b_o = \frac{\sum_h^L \frac{W_h^2\left(1-f_h\right)}{n_h}\cdot S_{xyh}}{\sum_h^L \frac{W_h^2\left(1-f_h\right)}{n_h}\cdot S_{yh}^2}$$

Pero como $\beta_h = \frac{S_{xyh}}{S_{yh}^2} \quad S_{xyh} = \beta_h S_{yh}^2$, se tiene $b_o = \frac{\sum_h^L \frac{W_h^2\left(1-f_h\right)}{n_h}\cdot S_{yh}^2\beta_h}{\sum_h^L \frac{W_h^2\left(1-f_h\right)}{n_h}\cdot S_{yh}^2}$

El valor b_o que minimiza la varianza del estimador combinado es entonces una medida ponderada de los coeficientes de regresión β_h, siendo las ponderaciones dadas por $\omega_h = \frac{W_h^2\left(1-f_h\right)}{n_h}\cdot S_{yh}^2$, de tal forma que se puede escribir $b_o = \frac{\sum_h^L \omega_h\beta_h}{\sum_h^L \omega_h} = \overline{\beta}_c$, pudiendo expresarse la varianza mínima como:

$$V_{min}\left(\overline{x}_{rgc}\right) = \sum_h^l W_h^2\,\frac{1-f_h}{n_h}\cdot\left(S_{xh}^2 + \overline{\beta}_c^{\,2}S_{yh}^2 - 2\overline{\beta}_c S_{xyh}\right)$$

que puede estimarse como:

$$\hat{V}_{min}\left(\overline{x}_{rgc}\right) = \sum_h^l W_h^2\,\frac{1-f_h}{n_h}\cdot\left(\hat{S}_{xh}^2 + \hat{\overline{\beta}}_c^{\,2}\hat{S}_{yh}^2 - 2\hat{\overline{\beta}}_c\hat{S}_{xyh}\right)$$

donde:

$$\hat{\overline{\beta}}_c = \frac{\sum_h^L \hat{\omega}_h\hat{\beta}_h}{\sum_h^L \hat{\omega}_h},\ \hat{\omega}_h = \frac{W_h^2\left(1-f_h\right)}{n_h}\cdot\hat{S}_{yh}^2 \text{ y } \hat{\beta}_h = \frac{\hat{S}_{xyh}}{\hat{S}_{yh}^2}.$$

Para estimar el total, el estimador combinado se forma como:

$$\hat{X}_{rgc} = \hat{X}_{st} + b_o\left(Y - \hat{Y}_{st}\right) = N\overline{x}_{st} + b_o\left(N\overline{Y} - N\overline{y}_{st}\right) = N\overline{x}_{rgc}$$

Su varianza puede entonces expresarse en función de la varianza para la estimación de la media de la siguiente forma:

$$V\left(\hat{X}_{rgc}\right) = V\left(N\overline{x}_{rgc}\right) = N^2V\left(\overline{x}_{rgc}\right) = N^2\sum_h^L \frac{W_h^2\left(1-f_h\right)}{n_h}\cdot\left(S_{xh}^2 + b_o^2S_{yh}^2 - 2b_oS_{xyh}\right)$$

pudiendo expresarse la varianza mínima como:

$$V_{min}\left(\hat{X}_{rgc}\right) = N^2 \sum_h^l W_h^2 \frac{1-f_h}{n_h} \cdot \left(S_{xh}^2 + \overline{\beta}_c^{\,2} S_{yh}^2 - 2\overline{\beta}_c S_{xyh}\right)$$

que puede estimarse como:

$$\hat{V}_{min}\left(\hat{X}_{rgc}\right) = N^2 \sum_h^l W_h^2 \frac{1-f_h}{n_h} \cdot \left(\hat{S}_{xh}^2 + \hat{\overline{\beta}}_c^{\,2} \hat{S}_{yh}^2 - 2\hat{\overline{\beta}}_c \hat{S}_{xyh}\right)$$

Muestreo con reposición

El valor b_o que minimiza la varianza del estimador combinado para la media es una media ponderada de los coeficientes de regresión β_h, siendo las ponderaciones dadas por $\omega_h = \frac{W_h^2}{n_h} \cdot \sigma_{yh}^2$, de tal forma que se puede escribir $b_o = \frac{\sum_h^L \omega_h \beta_h}{\sum_h^L \omega_h} = \overline{\beta}_c$, pudiendo expresarse la varianza mínima como:

$$V_{min}\left(\overline{x}_{rgc}\right) = \sum_h^l W_h^2 \frac{1}{n_h} \cdot \left(\sigma_{xh}^2 + \overline{\beta}_c^{\,2} \sigma_{yh}^2 - 2\overline{\beta}_c \sigma_{xyh}\right)$$

que puede estimarse como:

$$\hat{V}_{min}\left(\overline{x}_{rgc}\right) = \sum_h^l W_h^2 \frac{1}{n_h} \cdot \left(\hat{S}_{xh}^2 + \hat{\overline{\beta}}_c^{\,2} \hat{S}_{yh}^2 - 2\hat{\overline{\beta}}_c \hat{S}_{xyh}\right)$$

donde:

$$\hat{\overline{\beta}}_c = \frac{\sum_h^L \hat{\omega}_h \hat{\beta}_h}{\sum_h^L \hat{\omega}_h}, \; \hat{\omega}_h = \frac{W_h^2}{n_h} \cdot \hat{S}_{yh}^2 \;\text{ y }\; \hat{\beta}_h = \frac{\hat{S}_{xyh}}{\hat{S}_{yh}^2}.$$

Para estimar el total, la varianza puede entonces expresarse en función de la varianza para la estimación de la media de la siguiente forma:

$$V\left(\hat{X}_{rgc}\right) = V\left(N\overline{x}_{rgc}\right) = N^2 V\left(\overline{x}_{rgc}\right) = N^2 \sum_h^L \frac{W_h^2}{n_h} \cdot \left(\sigma_{xh}^2 + b_o^2 \sigma_{yh}^2 - 2b_o \sigma_{xyh}\right)$$

pudiendo expresarse la varianza mínima como:

$$V_{min}\left(\hat{X}_{rgc}\right) = N^2 \sum_h^l W_h^2 \frac{1}{n_h} \cdot \left(\sigma_{xh}^2 + \overline{\beta}_c^{\,2} \sigma_{yh}^2 - 2\overline{\beta}_c \sigma_{xyh}\right)$$

que puede estimarse como:

$$\hat{V}_{min}\left(\hat{X}_{rgc}\right) = N^2 \sum_h^l W_h^2 \frac{1}{n_h} \cdot \left(\hat{S}_{xh}^2 + \hat{\overline{\beta}}_c^{\,2} \hat{S}_{yh}^2 - 2\hat{\overline{\beta}}_c \hat{S}_{xyh}\right)$$

Comparación de precisiones en los stimadores de regresión separado y combinado

Vamos a comparar las varianzas mínimas de los estimadores de regresión separado y combinado. Tenemos:

$$V_{min}(\hat{\bar{X}}_{rg,c}) - V_{min}(\hat{\bar{X}}_{rg,s}) = \sum_{h=1}^{L} W_h^2 \frac{(1-f_h)}{n_h}\left[S_{Xh}^2 + \bar{\beta}_c^2 S_{Yh}^2 - 2\bar{\beta}_c S_{XYh}\right]$$

$$-\sum_{h=1}^{L} W_h^2 \frac{(1-f_h)}{n_h}\left[S_{Xh}^2 + \beta_h^2 S_{Yh}^2 - 2\beta_h S_{XYh}\right] = \sum_{h=1}^{L}\left[u_h\left(\bar{\beta}_c^2 - \beta_h^2\right) - 2u_h\left(\bar{\beta}_c - \beta_h\right)\beta_h\right]$$

$$= \sum_{h=1}^{L} u_h\left(\bar{\beta}_c - \beta_h\right)^2 \geq 0$$

Luego el estimador separado de regresión es más preciso que el combinado. Ambos tendrán igual varianza cuando $\bar{\beta}_c = \beta_h \ \forall h = 1,...,L$.

PROBLEMAS RESUELTOS

6.1. En un estudio para estimar el contenido total de azúcar de una carga de naranjas, se pesó una muestra de 10 naranjas, y se extrajo su jugo para pesar el contenido de azúcar. Se obtuvieron los siguientes resultados:

Naranja	Contenido de azúcar	Peso de la naranja
1	0,021	0,40
2	0,030	0,48
3	0,025	0,43
4	0,022	0,42
5	0,033	0,50
6	0,027	0,46
7	0,019	0,39
8	0,021	0,41
9	0,023	0,42
10	0,025	0,44

1) Sabiendo que el peso de todas las naranjas es 1800, estimar el contenido total de azúcar de las naranjas y su error de muestreo.

2) Estimar dichas varianzas y comparar la precisión de este tipo de muestreo con la del muestreo aleatorio simple. Seleccionar la muestra más precisa.

Como disponemos de información de una variable adicional muy correlacionada con la variable en estudio (ρ = 0,99), podemos realizar la estimación del contenido total de azúcar de las naranjas utilizando el estimador del total basado en la razón. Mediante el procedimiento *Estadística descriptiva* de la opción *Análisis de datos* del menú *Herramientas* (Figura 6-1), podemos calcular los estadísticos más relevantes relativos a la variable en estudio y a la variable adicional. La Figura 6-2 muestra los resultados.

Las fórmulas para los cálculos del estimador del total y de su error en la estimación por razón, $\hat{V}(\hat{X}_R) = N^2 \frac{1-f}{n}\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right)$, se muestran en la Figura 6-3, y los resultados en la Figura 6-4.

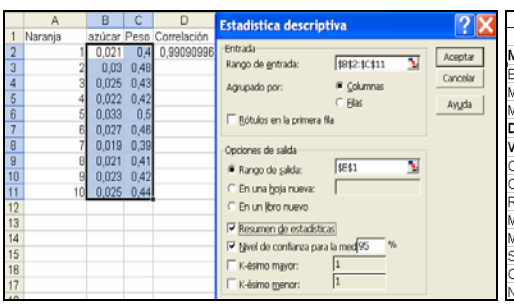


Figura 6-1

Columna1		Columna2	
Media	**0,0246**	**Media**	**0,435**
Error típico	0,001384036	Error típico	0,01118034
Mediana	0,024	Mediana	0,425
Moda	0,021	Moda	0,42
Desviación estándar	**0,004376706**	**Desviación estándar**	**0,035355339**
Varianza de la muestra	**1,91556E-05**	**Varianza de la muestra**	**0,00125**
Curtosis	-0,036561957	Curtosis	-0,367238095
Coeficiente de asimetría	0,784447173	Coeficiente de asimetría	0,707106781
Rango	0,014	Rango	0,11
Mínimo	0,019	Mínimo	0,39
Máximo	0,033	Máximo	0,5
Suma	0,246	Suma	4,35
Cuenta	10	Cuenta	10
Nivel de confianza(95,0%)	0,003130909	Nivel de confianza(95,0%)	0,025291705

Figura 6-2

	A	B	C	D	E	F	G	H
1	Naranja	azúcar	Peso	Correlación	Covarianza	Razón	Total estimado	Varianza
2	1	0,021	0,4	=COEF.DE.CORREL(B2:B11;C2:C11)	=COVAR(B2:B11;C2:C11)	=SUMA(azúcar)/SUMA(Peso)	=1800*Razón	=1800^2*(VAR(azúcar)+Razón^2*VAR(Peso)-2*Razón*Covarianza*10/9)/10
3	2	0,03	0,48					
4	3	0,025	0,43					
5	4	0,022	0,42					
6	5	0,033	0,5					
7	6	0,027	0,46					
8	7	0,019	0,39					
9	8	0,021	0,41					
10	9	0,023	0,42					
11	10	0,025	0,44					

Figura 6-3

	A	B	C	D	E	F	G	H
1	Naranja	azúcar	Peso	Correlación	Covarianza	Razón	Total estimado	Varianza
2	1	0,021	0,4	0,99090996	0,000138	0,05655172	**101,7931034**	**1,88265018**
3	2	0,03	0,48					
4	3	0,025	0,43					
5	4	0,022	0,42					
6	5	0,033	0,5					
7	6	0,027	0,46					
8	7	0,019	0,39					
9	8	0,021	0,41					
10	9	0,023	0,42					
11	10	0,025	0,44					

Figura 6-4

6.2. Consideramos una población de 500 individuos en la que está definida la característica bidimensional (*Xi,Yi*) que mide las ganancias mensuales en miles de euros de los varones (*X*) y las mujeres (*Y*) con título universitario superior. Una muestra aleatoria simple de tamaño 80 proporciona los siguientes datos:

$$\sum_{i=1}^{80} X_i = 420 \quad , \quad \sum_{i=1}^{80} Y_i = 190 \quad , \quad \sum_{i=1}^{80} X_i^2 = 2284 \quad , \quad \sum_{i=1}^{80} Y_i^2 = 512 \quad \text{y} \quad \sum_{i=1}^{80} X_i Y_i = 1045$$

1) Estimar la razón de las ganancias mensuales femeninas respecto de las masculinas, su sesgo y su error de muestreo. Estudiar la posible influencia del sesgo.

2) Se trata de estimar con y sin reposición la media y el total de las ganancias mensuales femeninas en la población utilizando la información adicional de la variable ganancia mensual masculina mediante un método de estimación indirecta. ¿Qué método indirecto sería el más adecuado? ¿Por qué? Realizar las estimaciones de las ganancias femeninas media y total mensuales mediante los métodos indirectos conocidos ordenándolos en precisión y sabiendo que la ganancia total masculina es 10000.

c) Cuantificar la ganancia en precisión respecto del muestreo aleatorio simple.

Tenemos:

$$\hat{S}_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{80} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{80} X_i\right)^2\right) = 1, \quad \hat{S}_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{80} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{80} Y_i\right)^2\right) = 0{,}768$$

$$\hat{S}_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{80} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{80} X_i\right)\left(\sum_{i=1}^{80} Y_i\right)\right) = 0{,}6012$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{80} X_i = 5{,}25 \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{80} Y_i = 2{,}375$$

Estimar la razón de las ganancias mensuales femeninas respecto de las masculinas es equivalente a estimar la razón de *Y* a *X*.

La razón *Y/X* se estima mediante $\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{y}{x} = 0{,}452$.

El sesgo del estimador de la razón anterior se estima mediante:

$$\hat{B}(\hat{R}) = \frac{(1-f)}{n\bar{x}^2}\left(\hat{R}\hat{S}_x^2 - \hat{S}_{XY}\right) = \frac{(1-80/500)}{80\cdot 5{,}25^2}\left(0{,}452\cdot 1 - 0{,}6012\right) = -0{,}0000568$$

El error de muestreo del estimador de la razón se estima mediante:

$$\hat{\sigma}(\hat{R}) = \sqrt{\frac{(1-f)}{n\bar{x}^2}\left(\hat{S}_y^2 + \hat{R}^2\hat{S}_x^2 - 2\hat{R}\hat{S}_{XY}\right)} = \sqrt{\frac{(1-80/500)}{80\cdot 5{,}25^2}\left(0{,}768 + 0{,}452^2\cdot 1 - 2\cdot 0{,}452\cdot 0{,}6012\right)} = 0{,}0128$$

Para ver si el sesgo del estimador de la razón es influyente hallamos el valor del sesgo relativo $\left|\frac{\hat{B}(\hat{R})}{\hat{\sigma}(\hat{R})}\right| = \frac{0{,}0000568}{0{,}0128} = 0{,}004 < 0{,}1$, por lo que el sesgo es despreciable.

Estimar la media y el total de las ganancias mensuales femeninas en la población es equivalente a estimar la media y el total de *Y*.

Para estudiar qué método de estimación indirecta es el más adecuado al estimar la media y el total de *Y* utilizamos la recta de regresión de la variable en estudio *Y* sobre la variable auxiliar *X*, cuya ecuación es:

$$y - \bar{y} = \frac{\hat{S}_{xy}}{\hat{S}_x^2}(x - \bar{x}) \Rightarrow y = 0{,}6012x - 0{,}78$$

Observamos que la recta de regresión de *Y* sobre *X* tiene una ordenada en el origen cercana a cero (comparada con los valores medios de *X* e *Y*), lo que indica que puede ser razonable la estimación indirecta de los parámetros poblacionales utilizando estimación basada en la razón. Además, el sesgo del estimador de la razón será pequeño (como ya hemos visto) porque la recta de regresión está próxima a pasar por el origen. Evidentemente, la estimación indirecta basada en regresión será la más apropiada, como ocurre siempre. Puede suceder que la estimación indirecta basada en la diferencia sea la menos apropiada ya que la pendiente de la recta de regresión no está claro que se aproxime a la unidad.

La utilización de métodos indirectos de estimación en todo el problema es apropiada, ya que el coeficiente de correlación $\hat{\rho} = \frac{\hat{S}_{xy}}{\hat{S}_x\hat{S}_y} \cong 0{,}7$ es alto.

Muestreo sin reposición

Comenzamos realizando estimaciones para la media y el total de la variable en estudio *Y* basadas en la razón de *Y* a la variable auxiliar *X* y a su vez calculamos también las varianzas de los estimadores.

$$\hat{\bar{Y}} = \hat{R}\bar{X} = \frac{\bar{y}}{\bar{x}}\bar{X} = 0{,}452 \cdot \frac{10000}{500} = 9{,}04 \quad \hat{Y} = \hat{R}X = \frac{\bar{y}}{\bar{x}}X = 0{,}452 \cdot 10000 = 4520$$

$$\hat{V}(\hat{\bar{Y}}) = \frac{(1-f)}{n}\left(\hat{S}_y^2 + \hat{R}^2\hat{S}_x^2 - 2\hat{R}\hat{S}_{XY}\right) = \frac{(1-\frac{80}{500})}{80}\left(0{,}768 + 0{,}452^2 \cdot 1 - 2 \cdot 0{,}452 \cdot 0{,}6012\right) = 0{,}0073$$

$$\hat{V}(\hat{Y}) = N^2\frac{(1-f)}{n}\left(\hat{S}_y^2 + \hat{R}^2\hat{S}_x^2 - 2\hat{R}\hat{S}_{XY}\right) = 500^2 \cdot 0{,}0073 = 1825$$

Ahora calculamos estimadores y varianzas basados en la regresión.

$$\hat{\bar{Y}}_{rg} = \bar{y} + b(\bar{X} - \bar{x}) = \bar{y} + \frac{\hat{S}_{xy}}{\hat{S}_x^2}(\bar{X} - \bar{x}) = 2{,}375 + \frac{0{,}6012}{1}\left(\frac{1000}{500} - 5{,}25\right) = 11{,}2427$$

$$\hat{Y}_{rg} = N\hat{\bar{Y}}_{rg} = 500 \cdot 11{,}2427 = 5621{,}35$$

$$\hat{V}_{min}(\hat{\bar{Y}}_{rg}) = \frac{(1-f)}{n}\hat{S}_y^2\left(1 - \hat{\rho}^2\right) = \frac{1-\frac{80}{500}}{80}0{,}768(1 - 0{,}7^2) = 0{,}004$$

$$\hat{V}_{min}(\hat{Y}_{rg}) = N^2\hat{V}_{min}(\hat{\bar{Y}}_{rg}) = 500^2 \cdot 0{,}004 = 1000$$

Ahora calculamos estimadores y varianzas basados en la diferencia.

$$\hat{\bar{Y}} = \hat{D} + \bar{X} = \bar{y} - \bar{x} + \bar{X} = 2{,}375 - 5{,}25 + \frac{10000}{500} = 17{,}125$$

$$\hat{Y} = \hat{D}_T + X = N(\bar{y} - \bar{x}) + N\bar{X} = N\hat{\bar{Y}} = 500 \cdot 17{,}125 = 8562{,}5$$

$$V(\hat{\bar{Y}}) = V(\hat{D} + \bar{X}) = V(\hat{D}) = \frac{(1-f)}{n}\left(\hat{S}_y^2 + \hat{S}_x^2 - \hat{S}_{XY}\right) = 0{,}009$$

$$V(\hat{Y}) = V(\hat{D}_T + X) = V(\hat{D}_T) = N^2V(\hat{D}) = 500^2 \cdot 0{,}009 = 2250$$

Ahora calculamos estimadores y varianzas para muestreo aleatorio simple.

$$\hat{\bar{Y}}_{as} = \bar{y} = 2{,}375 \qquad \hat{Y}_{as} = N\hat{\bar{Y}}_{as} = 500 \cdot 2{,}375 = 1187{,}5$$

$$\hat{V}(\hat{\bar{Y}}_{as}) = \frac{(1-f)}{n}\hat{S}_y^2 = \frac{1-\frac{80}{500}}{80}0{,}768 = 0{,}008$$

$$\hat{V}(\hat{Y}_{as}) = N^2\hat{V}(\hat{\bar{Y}}_{as}) = 500^2 \cdot 0{,}008 = 2000$$

Se observa que la menor varianza la presenta el estimador basado en la regresión, seguido del estimador basado en la razón, el estimador aleatorio simple y el estimador basado en la diferencia. Estos resultados coinciden con los especificados al principio del problema basados en la recta de regresión.

El estimador basado en la razón mejora al aleatorio simple si se cumple $\hat{\rho} > \frac{1}{2}\frac{\hat{C}_x}{\hat{C}_y}$

$$0{,}7 = \hat{\rho} > \frac{1}{2}\frac{\hat{C}_x}{\hat{C}_y} = \frac{\hat{S}_x}{\hat{S}_y}\hat{R} = \frac{1}{\sqrt{0{,}678}}0{,}452 = 0{,}5157$$

Por lo tanto, el muestreo basado en la razón es más preciso que el aleatorio simple. Ello implica que el muestreo basado en la regresión también es más preciso que el aleatorio simple. Sin embargo, ya hemos visto que el muestreo por diferencia es ligeramente menos preciso que el aleatorio simple.

La ganancia en precisión del estimador de regresión sobre el aleatorio simple es G = (0,008/0,004 – 1)100 = 100%.

La ganancia en precisión del estimador de razón sobre el aleatorio simple es G = (0,008/0,0073 – 1)100 = 9,5%.

La ganancia en precisión del estimador aleatorio simple sobre el de diferencia es G = (0,009/0,008 – 1)100 = 12,5%.

Muestreo con reposición

Las estimaciones de la media y total valen lo mismo que en muestreo sin reposición. Calculamos las estimaciones de las varianzas de los estimadores para estimación indirecta por razón.

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{n}\left(\hat{S}_y^2 + \hat{R}^2\hat{S}_x^2 - 2\hat{R}\hat{S}_{XY}\right) = 0{,}00869$$

$$\hat{V}(\hat{Y}) = N^2\frac{1}{n}\left(\hat{S}_y^2 + \hat{R}^2\hat{S}_x^2 - 2\hat{R}\hat{S}_{XY}\right) = 2172{,}5$$

Ahora estimamos varianzas basadas en la regresión.

$$\hat{V}_{min}(\hat{\bar{Y}}_{rg}) = \frac{1}{n}\hat{S}_y^2\left(1 - \hat{\rho}^2\right) = \frac{1}{80}0{,}768(1 - 0{,}7^2) = 0{,}00476$$

$$\hat{V}_{min}(\hat{Y}_{rg}) = N^2\hat{V}_{min}(\hat{\bar{Y}}_{rg}) = 500^2 \cdot 0{,}00476 = 11900$$

Ahora estimamos varianzas basadas en la diferencia.

$$V(\hat{\bar{Y}}) = V(\hat{D} + \bar{X}) = V(\hat{D}) = \frac{1}{n}\left(\hat{S}_y^2 + \hat{S}_x^2 - \hat{S}_{XY}\right) = 0{,}0107$$

$$V(\hat{Y}) = V(\hat{D}_T + X) = V(\hat{D}_T) = N^2V(\hat{D}) = 500^2 \cdot 0{,}0107 = 2675$$

Ahora estimamos varianzas para muestreo aleatorio simple.

$$\hat{V}(\hat{\bar{Y}}_{as}) = \frac{1}{n}\hat{S}_y^2 = \frac{1}{80}0{,}768 = 0{,}0096$$

$$\hat{V}(\hat{Y}_{as}) = N^2\hat{V}(\hat{\bar{Y}}_{as}) = 500^2 \cdot 0{,}0096 = 2400$$

Se observa que la menor varianza la presenta el estimador basado en la regresión, seguido del estimador basado en la razón, el estimador aleatorio simple y el estimador basado en la diferencia. Estos resultados presentan varianzas mayores que en el caso de sin reposición para todos los estimadores, ya que el muestreo con reposición es menos preciso que el muestreo sin reposición.

La ganancia en precisión del estimador de regresión sobre el aleatorio simple es G = (0,0096/0,00476 – 1)100 = 101,6%.

La ganancia en precisión del estimador de razón sobre el aleatorio simple es G = (0,0096/0,00869 – 1)100 = 10,47%.

La ganancia en precisión del estimador aleatorio simple sobre el de diferencia es G = (0,0107/0,0096 – 1)100 = 11,45%.

Se observa que la utilización del método indirecto de estimación basado en la regresión mejora fuertemente la estimación aleatoria simple, y que la utilización del método indirecto de estimación basado en la razón mejora levemente la estimación aleatoria simple. Las ganancias en precisión se han acentuado levemente respecto del muestreo sin reposición. El método indirecto de la diferencia es ligeramente peor que el aleatorio simple; sin embargo, la ganancia en precisión del aleatorio simple sobre la estimación por diferencia disminuye al considerar reposición.

6.3. De los $N = 750$ trabajadores de una fábrica se conoce que el número medio de días anuales de ausencia del trabajo sin justificar para las mujeres (variable X) es 10 y para los hombres (variable Y) es 8. Se sabe que el error cometido al cuantificar la media de la variable X es 2500 y que la razón de la covarianza de X e Y a la varianza de X es 0,6. Determinar a partir de qué tamaño muestral el sesgo del estimador de la razón Y/X es despreciable utilizando muestreo sin y con reposición. ¿Qué método de estimación indirecta sería el más adecuado a utilizar sobre muestras de esta población?

Determinar a partir de qué tamaño muestral el sesgo del estimador de la razón Y/X es despreciable utilizando muestreo sin y con reposición. ¿Qué método de estimación indirecta sería el más adecuado a utilizar sobre muestras de esta población?

El enunciado del problema nos da como datos:

$$\overline{X} = 10\,, \quad \overline{Y} = 8\,, \quad \sigma_x^2 = 2500 \quad \text{y} \quad \frac{\sigma_{xy}}{\sigma_x^2} = 0{,}6$$

De la condición de que el sesgo relativo $\left|\frac{B(\hat{R})}{\sigma(\hat{R})}\right|$ sea menor que un décimo se obtiene que $n \geq \dfrac{N \cdot 100 \cdot S_x^2}{N\overline{X}^2 + 100 S_x^2} = \dfrac{750 \cdot 100 \cdot \frac{750}{749} 2500}{750 \cdot 10^2 + 100 \frac{750}{749} 2500} = 577$.

En caso de muestreo con reposición la misma condición de sesgo relativo menor que un décimo nos lleva a $n \geq 100\frac{\sigma_x^2}{\overline{X}^2} = 100\frac{2500}{100} = 2500$, que sobrepasa el tamaño poblacional (con los datos del problema nunca podría ser el sesgo despreciable).

La recta de regresión de Y sobre X tiene de ecuación $y - \overline{y} = \frac{\hat{S}_{xy}}{\hat{S}_x^2}(x - \overline{x})$ $\Rightarrow y - 8 = 0{,}6(x - 10) \Rightarrow y = 0{,}6x + 2$, lo que indica que la estimación por razón podría ser adecuada al no ser demasiado grande la ordenada en el origen. La estimación por regresión siempre es el método más adecuado. La pendiente de la recta no es unitaria, con lo que no es muy apropiada la estimación por diferencia.

6.4. Para estudiar el grado medio de implantación de un determinado cultivo en una región se obtuvo una muestra de 100 fincas para las que se midió la superficie dedicada al cultivo en estudio (variable *X*) y su superficie total (variable *Y*), obteniéndose los datos que se presentan en la tabla adjunta. Se pide:

1º) A la vista de la información, justificar si será adecuado el uso de los métodos indirectos de muestreo respecto del muestreo aleatorio simple y estudiar qué métodos serán los más adecuados expresándolos por orden de preferencia. Hallar los errores relativos de muestreo para los diferentes métodos cuantificando sesgos y ganancias en precisión y razonando adecuadamente los resultados. Contrastar también los resultados obtenidos considerando muestreo con reposición y sin reposición.

2º) Dada la estructura de las fincas se consideró conveniente realizar una estratificación según la variable superficie total de la fincas. Se consideraron dos estratos relativos a fincas de superficie total superior a una hectárea y a fincas de superficie total menor o igual que una hectárea. Los datos obtenidos también se presentan en la tabla adjunta. A la vista de esta información, justificar si serán adecuados los métodos de estimación indirecta con estratificación y cuál de entre ellos puede resultar mejor. Hallar los errores relativos de muestreo para los diferentes métodos de estimación con muestreo estratificado cuantificando sesgos y ganancias en precisión y razonando adecuadamente los resultados. Contrastar también los resultados obtenidos considerando muestreo con reposición y sin reposición.

Estratos	*Superficie de las fincas*	N_h	$\hat{S}_{yh}^{\ 2}$	$\hat{S}_{xh}^{\ 2}$	$\hat{\rho}_{xyh}$	$\overline{y}_h$	$\overline{x}_h$	n_h
1	$0-1Ht$	1580	2055	312	0.62	82.5	19.4	70
2	$>1Ht$	430	7357	922	0.3	244.8	51.6	30
Población			7619	620	0.67			

Se trata de estimar con y sin reposición la media y el total de *Y* utilizando la información adicional de la variable *X* mediante un método de estimación indirecta. ¿Qué método indirecto sería el más adecuado? ¿Por qué? Realizar las estimaciones de media y total mediante los métodos indirectos conocidos ordenándolos en precisión y sabiendo que el total de *X* es 10000.

Tenemos como dato que $\hat{\rho} = \frac{\hat{S}_{xy}}{\hat{S}_x \hat{S}_y} = 0{,}67$, por lo que la utilización de métodos indirectos de estimación en todo el problema es apropiada, ya que el coeficiente de correlación estimado es alto.

Para estudiar qué método de estimación indirecta es el más adecuado al estimar la superficie dedicada al cultivo (variable *X*) en las fincas utilizamos la recta de regresión de la variable en estudio *X* sobre la variable auxiliar *Y* superficie total de las fincas, cuya ecuación es:

$$x - \bar{x} = \frac{\hat{S}_{xy}}{\hat{S}_y^2}(y - \bar{y}) \Rightarrow x - 26{,}3 = \frac{1453}{7619}(y - 117{,}28) \Rightarrow x = 0{,}19y + 4$$

$$\bar{x} = \sum_{h=1}^{2} W_h \bar{x}_h = \frac{N_1}{N}\bar{x}_1 + \frac{N_2}{N}\bar{x}_2 = \frac{1580}{2010}19{,}4 + \frac{430}{2010}51{,}63 = 26{,}3$$

$$\bar{y} = \sum_{h=1}^{2} W_h \bar{y}_h = \frac{N_1}{N}\bar{y}_1 + \frac{N_2}{N}\bar{y}_2 = \frac{1580}{2010}82{,}56 + \frac{430}{2010}244{,}85 = 117{,}28$$

$$\hat{S}_{xy} = \hat{\rho}_{xy}\hat{S}_x\hat{S}_y \Rightarrow 0{,}67\sqrt{620}\sqrt{7619} = 1453 \qquad \hat{R} = \frac{\bar{x}}{\bar{y}} = \frac{26{,}30}{117{,}28} = 0{,}224$$

Observamos que la recta de regresión de *X* sobre *Y* tiene una ordenada en el origen que no se anula, pero es pequeña (comparada con los valores medios de *X* e *Y*), lo que indica que puede ser razonable la estimación indirecta de los parámetros poblacionales utilizando estimación basada en la razón. Además el sesgo del estimador de la razón será pequeño porque la recta de regresión está próxima a pasar por el origen. Evidentemente, la estimación indirecta basada en regresión será la más apropiada, como ocurre siempre. La estimación indirecta basada en la diferencia será la menos apropiada, ya que la pendiente de la recta de regresión no se aproxima a la unidad.

El estimador basado en la razón mejora al aleatorio simple si se cumple $\hat{\rho} > \frac{1}{2}\frac{\hat{C}_y}{\hat{C}_x}$

$$0{,}67 = \hat{\rho} > \frac{1}{2}\frac{\hat{C}_y}{\hat{C}_x} = \frac{\hat{S}_y}{\hat{S}_x}\hat{R} = \frac{1}{2}\frac{\sqrt{7619}}{\sqrt{620}}\frac{26{,}30}{117{,}28} = 0{,}393$$

Por lo tanto, el muestreo basado en la razón es más preciso que el aleatorio simple. Ello implica que el muestreo basado en la regresión también es más preciso que el aleatorio simple. Sin embargo, ya hemos razonado que el muestreo por diferencia probablemente será menos preciso que el aleatorio simple, y, por tanto, también será menos preciso que la estimación por razón y regresión. Vamos a realizar los cálculos de varianzas.

Muestreo sin reposición

Comenzamos hallando el error para la estimación de la media (grado medio de implantación del cultivo medido a través de la superficie dedicada al cultivo) de la variable en estudio *X* basada en la razón de *X* a la variable auxiliar *Y*.

$$\hat{V}(\hat{\bar{X}}) = \frac{(1-f)}{n}\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{XY}\right) = \frac{(1-\frac{100}{2010})}{100}\left(620 + 0{,}224^2 \cdot 7619 - 2 \cdot 0{,}224 \cdot 1453\right) = 3{,}335$$

Ahora estimamos el error del estimador de la media basado en la regresión.

$$\hat{V}_{\min}(\hat{\bar{X}}_{rg}) = \frac{(1-f)}{n}\hat{S}_x^2\left(1-\hat{\rho}^2\right) = \frac{1-\frac{100}{2010}}{100}620(1-0{,}67^2) = 3{,}24$$

Ahora estimamos el error del estimador de la media basado en la diferencia.

$$V(\hat{\bar{X}}) = V(\hat{D}+\bar{Y}) = V(\hat{D}) = \frac{(1-f)}{n}\left(\hat{S}_x^2 + \hat{S}_y^2 - \hat{S}_{XY}\right) = \frac{1-\frac{100}{2010}}{100}(620+7619-1453) = 64{,}4$$

Ahora estimamos el error del estimador de la media en el aleatorio simple.

$$\hat{V}(\hat{\bar{X}}_{as}) = \frac{(1-f)}{n}\hat{S}_x^2 = \frac{1-\frac{100}{2010}}{100}620 = 5{,}89$$

Se observa que la menor varianza la presenta el estimador basado en la regresión, seguido del estimador basado en la razón, el estimador aleatorio simple y el estimador basado en la diferencia. Estos resultados coinciden con los especificados al principio del problema basados en la recta de regresión.

La ganancia en precisión del estimador de regresión sobre el aleatorio simple es G = (5,89/3,24 – 1)100 = 81,8%.

La ganancia en precisión del estimador de razón sobre el aleatorio simple es G = (5,89/3,335 – 1)100 = 76,6%.

La ganancia en precisión del estimador de regresión sobre el de razón es G = (3,335/3,24 – 1)100 = 2,9%.

En cuanto a la estimación del sesgo de estimador de la razón tenemos:

$$\hat{B}(\hat{R}) = \frac{(1-f)}{n\bar{y}^2}\left(\hat{R}\hat{S}_y^2 - \hat{S}_{XY}\right) = \frac{1-\frac{100}{2010}}{100 \cdot 117{,}28}\left(0{,}224 \cdot 7619 - 1453\right) = 0{,}02$$

Este sesgo resulta despreciable porque 0,02/3,335 = 0,006 < 1/10.

Muestreo con reposición

Comenzamos estimando el error del estimador de la media de la variable en estudio *X* basado en la razón de *X* a la variable auxiliar *Y*.

$$\hat{V}(\hat{\bar{X}}) = \frac{1}{n}\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{XY}\right) = \frac{1}{100}\left(620 + 0{,}224^2 \cdot 7619 - 2 \cdot 0{,}224 \cdot 1453\right) = 3{,}51$$

Ahora estimamos el error del estimador de la media basado en regresión.

$$\hat{V}_{min}(\hat{\bar{X}}_{rg}) = \frac{1}{n}\hat{S}_x^2\left(1-\hat{\rho}^2\right) = \frac{1}{100}620(1-0{,}67^2) = 3{,}41$$

Ahora estimamos el error del estimador de la media basado en diferencia.

$$V(\hat{\bar{X}}) = V(\hat{D}+\bar{Y}) = V(\hat{D}) = \frac{1}{n}\left(\hat{S}_x^2+\hat{S}_y^2-\hat{S}_{XY}\right) = \frac{1}{100}(620+7619-1453) = 67{,}78$$

Ahora estimamos el error del estimador de la media en el aleatorio simple.

$$\hat{V}(\hat{\bar{X}}_{as}) = \frac{\hat{S}_x^2}{n} = \frac{620}{100} = 6{,}2$$

Se observa que la menor varianza la presenta el estimador basado en la regresión, seguido del estimador basado en la razón, el estimador aleatorio simple y el estimador basado en la diferencia. Estos resultados son superiores a los correspondientes a muestreo sin reposición debido a que el muestreo con reposición es menos preciso.

El sesgo del estimador de la razón se estima mediante:

$$\hat{B}(\hat{R}) = \frac{1}{n\bar{y}^2}\left(\hat{R}\hat{S}_y^2-\hat{S}_{XY}\right) = \frac{(1-600/1500)}{600\cdot 5{,}58^2}\left(2\cdot 7-3{,}75\right) = 0{,}0005$$

Consideramos ahora la estratificación en dos estratos según la superficie total de las fincas, y vamos a considerar las estimaciones separada y combinada para la media en razón y regresión para calcular sus errores de muestreo y sus sesgos.

Comenzaremos determinando valores necesarios en todos los cálculos posteriores, como son: $W_1 = 1580/2010 = 0{,}786$, $W_2 = 430/2010 = 0{,}214$, $f_1 = 70/100 = 0{,}7$, $f_2 = 30/100 = 0{,}3$, $\hat{R}_1 = 19{,}40/82{,}56 = 0{,}235$, $\hat{R}_2 = 51{,}63/244{,}85 = 0{,}21$, $\hat{S}_{xy1} = \hat{\rho}_{xy1}\hat{S}_x\hat{S}_y = 496{,}4$ y $\hat{S}_{xy2} = \hat{\rho}_{xy2}\hat{S}_x\hat{S}_y = 781{,}3$.

Estimador combinado de la razón

La *estimación combinada de la varianza del estimador de la media para muestreo sin reposición* será $\hat{V}(\hat{\bar{X}}_{RC}) = \sum_h^L \frac{W_h^2(1-f_h)}{n_h}\left(\hat{S}_{xh}^2+\hat{R}^2\hat{S}_{yh}^2-2\hat{R}\,\hat{S}_{xyh}\right) = 1{,}51593$.

El *sesgo del estimador combinado para la media puede estimarse como*:

$\hat{B}(\hat{\bar{X}}_{RC}) = \sum_h^L \frac{W_h^2(1-f_h)}{n_h\bar{Y}}\left(\hat{R}\hat{S}_{Yh}^2-\hat{S}_{XYh}\right) = 0{,}83/\bar{Y}$. Las operaciones a realizar son:

Pero $\bar{Y}$ se estima por $\bar{y} = 117{,}2 \Rightarrow \hat{B}(\hat{\bar{X}}_{RC}) = 0{,}83/117{,}2 = 0{,}007$.

La *estimación de la varianza de la media para muestreo con reposición* será:

$$\hat{V}(\hat{\bar{X}}_{RC}) = \sum_h^L \frac{W_h^2}{n_h}\left(\hat{S}_{xh}^2+\hat{R}^2\hat{S}_{yh}^2-2\hat{R}\hat{S}_{xyh}\right) = 3{,}1375\,.$$

Para muestreo con reposición el sesgo puede estimarse como:

$$\hat{B}(\hat{\bar{X}}_{RC}) = \sum_{h}^{L} \frac{W_h^2}{n_h \bar{Y}} \left(\hat{R} \hat{S}_{Yh}^2 - \hat{S}_{XYh} \right) = 1.00456/\bar{Y}$$

Pero $\bar{Y}$ se estima por $\bar{y} = 117{,}2 \Rightarrow \hat{B}(\hat{\bar{X}}_{RC}) = 1{,}00456/117{,}2 = 0{,}0085$.

Estimador separado de la razón

La *estimación de la varianza del estimador de la media para muestreo sin reposición* será:

$$\hat{V}(\hat{\bar{X}}_{RS}) = \sum_{h}^{L} \frac{W_h^2 (1 - f_h)}{n_h} \left(\hat{S}_{xh}^2 + \hat{R}_h^2 \hat{S}_{yh}^2 - 2 \hat{R}_h \hat{S}_{xyh} \right) = 1{,}49.$$

El valor del *sesgo del estimador simple o separado* sin reposición *puede estimarse como*: $\hat{B}(\hat{\bar{X}}_{RS}) = \sum_{h}^{L} \frac{W_h (1 - f_h)}{n_h \bar{Y}_h} \left(\hat{R}_h \hat{S}_{Yh}^2 - \hat{S}_{XYh} \right) = 0{,}0029$. $\bar{Y}_1$ e $\bar{Y}_2$ se estimarán mediante $\bar{y}_1$ e $\bar{y}_2$ respectivamente. Los cálculos a realizar serían:

La *varianza del estimador separado de la media para muestreo con reposición* puede estimarse como $\hat{V}(\hat{X}_{RS}) = \sum_{h}^{L} \frac{W_h^2}{n_h} \left(\hat{S}_{xh}^2 + \hat{R}_h^2 \hat{S}_{yh}^2 - 2 \hat{R}_h \hat{S}_{xyh} \right) = 3{,}09792$.

Para muestreo con reposición la expresión del sesgo puede estimarse como: $\hat{B}(\hat{X}_{RS}) = \sum_{h}^{L} \frac{W_h}{n_h \bar{Y}_h} \left(\hat{R}_h \hat{S}_{Yh}^2 - \hat{S}_{XYh} \right) = 0{,}0033$.

Estimador combinado en regresión

La estimación de la varianza mínima del estimador de la media viene expresada en *muestreo sin reposición* por la expresión:

$$\hat{V}_{min}\left(\bar{x}_{rgc} \right) = \sum_{h}^{l} W_h^2 \frac{1 - f_h}{n_h} \cdot \left(\hat{S}_{xh}^2 + \hat{\bar{\beta}}_c^2 \hat{S}_{yh}^2 - 2 \hat{\bar{\beta}}_c \hat{S}_{xyh} \right)$$

donde:

$$\hat{\bar{\beta}}_c = \frac{\sum_{h}^{L} \hat{\omega}_h \hat{\beta}_h}{\sum_{h}^{L} \hat{\omega}_h} 0{,}16155 \quad \text{con } \hat{\omega}_h = \frac{W_h^2 \left(1 - f_h \right)}{n_h} \cdot \hat{S}_{yh}^2 \text{ y } \hat{\beta}_h = \frac{\hat{S}_{xyh}}{\hat{S}_{yh}^2}.$$

Calculado $\hat{\bar{\beta}}_c$ ya podemos hallar el valor de la varianza mínima mediante:

$$\hat{V}_{min}\left(\bar{x}_{rgc} \right) = \sum_{h}^{l} W_h^2 \frac{1 - f_h}{n_h} \cdot \left(\hat{S}_{xh}^2 + \hat{\bar{\beta}}_c^2 \hat{S}_{yh}^2 - 2 \hat{\bar{\beta}}_c \hat{S}_{xyh} \right) = 1{,}46407.$$

La estimación de la varianza mínima del estimador de la media viene expresada en *muestreo con reposición* por la expresión:

$$\hat{V}_{min}\left(\bar{x}_{rgc}\right) = \sum_{h}^{l} W_h^2 \frac{1}{n_h} \cdot \left(\hat{S}_{xh}^2 + \hat{\bar{\beta}}_c^2 \hat{S}_{yh}^2 - 2\hat{\bar{\beta}}_c \hat{S}_{xyh} \right)$$

donde:

$$\hat{\bar{\beta}}_c = \frac{\sum_{h}^{L} \hat{\omega}_h \hat{\beta}_h}{\sum_{h}^{L} \hat{\omega}_h} = 0{,}18977 \quad \text{con} \quad \hat{\omega}_h = \frac{W_h^2}{n_h} \cdot \hat{S}_{yh}^2 \quad \text{y} \quad \hat{\beta}_h = \frac{\hat{S}_{xyh}}{\hat{S}_{yh}^2}.$$

Calculado $\hat{\bar{\beta}}_c$ ya podemos hallar el valor de la varianza mínima mediante:

$$\hat{V}_{min}\left(\bar{x}_{rgc}\right) = \sum_{h}^{l} W_h^2 \frac{1}{n_h} \cdot \left(\hat{S}_{xh}^2 + \hat{\bar{\beta}}_c^2 \hat{S}_{yh}^2 - 2\hat{\bar{\beta}}_c \hat{S}_{xyh} \right) = 3{,}10321.$$

Estimador separado en regresión

La estimación de la varianza mínima del estimador de la media viene expresada en *muestreo sin reposición* por la expresión:

$$\hat{V}_{\min}\left(\bar{x}_{rgst}\right) = \sum_{h}^{L} W_h^2 \frac{1-f_h}{n_h} (\hat{S}_{Xh}^2 + \hat{\beta}_h \hat{S}_{Xh}^2 - 2\hat{\beta}_h \hat{S}_{XYh}) = \sum_{h}^{L} W_h^2 \frac{1-f_h}{n_h} \hat{S}_{xh}^2 (1 - \hat{\rho}^2{}_{xyh}) = 1{,}40509$$

La estimación de la varianza mínima del estimador de la media viene expresada en *muestreo con reposición* por la expresión:

$$\hat{V}_{min}\left(\bar{x}_{rgst}\right) = \sum_{h}^{L} W_h^2 \frac{1}{n_h} (\hat{S}_{Xh}^2 + \hat{\beta}_h \hat{S}_{Xh}^2 - 2\hat{\beta}_h \hat{S}_{XYh}) = \sum_{h}^{L} W_h^2 \frac{1}{n_h} \hat{S}_{xh}^2 (1 - \hat{\rho}^2{}_{xyh}) = 2{,}97591$$

Resumiendo resultados tenemos:

- *ESTRATIFICADO*
  - *RAZÓN*
    - *SEPARADA*
      - *SIN REPOSICIÓN* → 1,49
      - *CON REPOSICIÓN* → 3,09792
    - *COMBINADA*
      - *SIN REPOSICIÓN* → 1,51593
      - *CON REPOSICIÓN* → 3,1375
  - *REGRESIÓN*
    - *SEPARADA*
      - *SIN REPOSICIÓN* → 1,40509
      - *CON REPOSICIÓN* → 2,97591
    - *COMBINADA*
      - *SIN REPOSICIÓN* → 1,46407
      - *CON REPOSICIÓN* → 3,10321
- *SIN ESTRATIFICAR*
  - *RAZÓN*
    - *SIN REPOSICIÓN* → 3,335
    - *CON REPOSICIÓN* → 3,51
  - *REGRESIÓN*
    - *SIN REPOSICIÓN* → 3,24
    - *CON REPOSICIÓN* → 3,41

6.5. En una determinada comunidad se intenta estudiar el cambio relativo en el valor catastral de los bienes inmuebles en los dos últimos años. Se selecciona una muestra irrestricta aleatoria de $n = 20$ inmuebles de entre los N = 1000 de la comunidad. De los registros fiscales se obtiene el valor catastral para este año (X) o valor actual y el valor correspondiente de hace dos años (Y) o valor calculado, de cada una de las $n = 20$ casas incluidas en la muestra. Se desea estimar R, el cambio relativo en el valor catastral para los $N = 1000$ inmuebles de la comunidad, usando la información contenida en la muestra.

Casa	Valor calculadc	Valor actual			
	y_i	x_i	y_i^2	x_i^2	$x_i y_i$
1	6,7	7,1	44,89	50,41	47,57
2	8,2	8,4	67,24	70,56	68,88
3	7,9	8,2	62,41	67,24	74,78
4	6,4	6,9	40,96	47,61	44,16
5	8,3	8,4	68,89	70,56	69,72
6	7,2	7,9	51,84	62,41	56,88
7	6	6,5	36	42,24	39
8	7,4	7,6	54,76	57,76	56,24
9	8,1	8,9	65,61	79,21	72,09
10	9,3	9,9	86,49	98,01	92,07
11	8,2	9,1	67,24	82,81	74,62
12	6,8	7,3	46,24	53,29	49,64
13	7,4	7,8	54,76	60,84	57,72
14	7,5	8,3	56,25	68,89	62,25
15	8,3	8,9	68,89	79,21	73,87
16	9,1	9,6	82,81	92,16	87,36
17	8,6	8,7	73,96	75,69	74,82
18	7,9	8,8	62,41	77,44	69,52
19	6,3	7	39,69	49	44,1
20	8,9	9,4	79,21	88,36	83,66
Total	154,5	164,7	1210,55	1373,71	1288,95

La estimación del cambio relativo R en el valor catastral desde hace dos años se obtiene mediante el estimador de razón siguiente:

$$\hat{R} = \frac{\hat{X}}{\hat{Y}} = \frac{\bar{x}}{\bar{y}} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} = \frac{164,7}{154,5} = 1,07$$

$$\hat{V}\left(\hat{R}\right) = \frac{1-f}{\bar{Y}^2 n} \cdot \left(\hat{S}_x^2 + \hat{R}^2 \hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right) = \frac{1-f}{\bar{Y}^2 n(n-1)} \cdot \left[\sum_i^n X_i^2 + \hat{R}^2 \sum_i^n Y_i^2 - 2\hat{R}\sum_i^n X_i Y_i\right]$$

Como $\bar{Y}^2$ no se conoce, se estima mediante $\bar{y}^2$. Tenemos:

$$\hat{V}\left(\hat{R}\right) = \frac{1 - 20/100}{(154,5/20)^2 (20)(19)} \cdot \left[1373,71 + 1,07^2 (1210,55) - 2(1,07)1288,95\right] = 0,0001.$$

Por tanto, el error de muestreo es $\hat{\sigma}\left(\hat{R}\right)=\sqrt{\hat{V}\left(\hat{R}\right)}=\sqrt{0{,}0001}=0{,}01$.

El error relativo de muestreo será:

$$\hat{C}v\left(\hat{R}\right)=\frac{\hat{\sigma}\left(\hat{R}\right)}{\hat{R}}=\frac{0{,}01}{1{,}07}=0{,}0093\approx 1\%$$

Como el cambio relativo del valor catastral de los inmuebles se ha estimado en 1,07, la subida en los dos últimos años se estima que es del 7%, con un error del 1%.

6.6. Una compañía desea estimar la cantidad promedio de dinero μ_x pagado a los empleados por gastos médicos durante los tres primeros meses del año en curso. Los resultados del promedio por trimestres μ_y están disponibles en los informes fiscales del año anterior. Una muestra aleatoria de 100 registros de empleados se seleccionó de una población de 1000 empleados. Los resultados de la muestra se resumen a continuación:

$$n = 100,\ N = 1000$$

Total para el trimestre actual: $\sum_{i=1}^{100} x_i = 1750$

Total para el trimestre correspondiente del año anterior: $\sum_{i=1}^{100} y_i = 1200$

Total poblacional para el trimestre correspondiente del año anterior $\sum_{i=1}^{1000} y_i = 12500$

$$\sum_{i=1}^{100} x_i^2 = 31650 \qquad \sum_{i=1}^{100} y_i^2 = 15620 \qquad \sum_{i=1}^{100} y_i x_i = 22059{,}35.$$

Usar los datos para estimar μ_y y establecer un límite para el error de estimación.

Como tenemos información de una variable auxiliar Y, la utilizaremos para realizar una estimación indirecta de X basada en la razón de X a Y. Tenemos:

$$\hat{\bar{X}}_R = \bar{x}_R = \frac{\bar{x}}{\bar{y}}\bar{Y} = \hat{R}\bar{Y} = \frac{\sum_{i=1}^{100} x_i}{\sum_{i=1}^{100} y_i}\frac{12500}{1000} = \frac{1750}{1200}\frac{12500}{1000} = 18{,}23$$

El error para la estimación anterior se estima mediante:

$$\hat{V}(\hat{\bar{X}}_R) = \frac{1-f}{n}\left(\hat{S}_x^2 + \hat{R}^2\hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right) = \frac{1-f}{n(n-1)}\left[\sum_i^n X_i^2 + \hat{R}^2\sum_i^n Y_i^2 - 2\hat{R}\sum_i^n X_i Y_i\right]$$

$$\hat{V}(\bar{x}_R)= \frac{1-\frac{100}{1000}}{100(100-1)}\left[31650+\left(\frac{1750}{1200}\right)15620-2\frac{1750}{1200}22059{,}35\right]=0{,}0441$$

Un límite para el error de estimación al 95% será $2\sqrt{\hat{V}(\bar{x}_R)}=0{,}42$.

Hemos estimado que la cantidad promedio de dinero pagado a los empleados por gastos médicos es 18,23 unidades monetarias y tenemos una confianza alta de que el error cometido no supera las 0,42 unidades monetarias.

6.7. Se trata de realizar un estudio sobre las granjas de cerdos en una determinada comarca analizando una muestra obtenida en 10 municipios. Para ello se estratifica la comarca en dos zonas, una de secano y otra de regadío. En cada zona se mide el número de granjas existente (variable *X*) y el número de cerdos (variable *Y*) por municipios muestrales. Se obtienen los siguientes datos:

Zona	*Secano*				*Regadío*					
Fracción de muestreo	10%				20%					
Número de granjas	71				182					
Municipio muestral	1	2	3	4	1	2	3	4	5	6
X	1	3	2	1	5	8	6	7	6	5
Y	10	25	22	11	55	90	61	77	66	51

Se pide:

1) Estimar de la forma más eficiente posible el número total de cerdos y el promedio de cerdos por granja en el supuesto de que la selección de los municipios de la muestra haya sido con reposición. Razonar la elección de los estimadores.

2) Hallar el tamaño muestral necesario para cometer un error del 10% al estimar el número total de cerdos mediante muestreo estratificado con afijación proporcional al número de granjas existentes en cada municipio y realizar la afijación.

Sean:

X_{ih}= Número de granjas de cerdos existentes en el municipio muestral i-ésimo del estrato h-ésimo.

Y_{ih}= Número de cerdos existentes en el conjunto de explotaciones ganaderas del municipio muestral *i*-ésimo del estrato *h*-ésimo.

Tenemos:

$$f_1=\frac{n_1}{N_1}\Rightarrow 0{,}1=\frac{4}{N_1}\Rightarrow N_1=40 \qquad f_2=\frac{n_2}{N_2}\Rightarrow 0{,}2=\frac{6}{N_2}\Rightarrow N_2=30$$

Vamos a estimar el número total de cerdos en las granjas y sus errores absoluto y relativo de muestreo mediante muestreo estratificado como sigue:

$$\hat{Y} = \sum_{h=1}^{2} N_h \bar{y}_h = N_1 \bar{y}_1 + N_2 \bar{y}_2 = 40\frac{10+25+22+11}{4} + 30\frac{55+90+61+77+66+51}{6} = 2780$$

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{2} N_h^2 \frac{\hat{S}_{yh}^2}{n_h} = 40^2 \frac{\hat{S}_{y1}^2}{4} + 30^2 \frac{\hat{S}_{y2}^2}{6} = 40^2 \frac{7.61}{4} + 30^2 \frac{30{,}15}{6} = 7566{,}5$$

$$\hat{S}_{yh}^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(Y_{hi} - \bar{y}_h\right)^2 \Rightarrow \begin{cases} \hat{S}_{y1}^2 = 7{,}61 \\ \hat{S}_{y2}^2 = 30{,}15 \end{cases} \quad \hat{\sigma}(\hat{Y}) = \sqrt{\hat{V}(\hat{Y})} = \sqrt{7566{,}5} = 87$$

$$\hat{Cv}(\hat{Y}) = \frac{\hat{\sigma}(\hat{Y})}{\hat{Y}} = \frac{87}{2780} = \frac{\sqrt{6357{,}67}}{2780} = 0{,}0312 \quad (3{,}12\%)$$

Para estimar el promedio de cerdos por explotación ganadera utilizamos el estimador de razón de Y a X (también puede usarse razón separada o combinada).

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{2} N_h \bar{y}_h}{\sum_{h=1}^{2} N_h \bar{x}_h} = \frac{2780}{40\frac{1+2+3+1}{4} + 30\frac{5+8+6+7+6+5}{6}} = \frac{2780}{255} = 10{,}9$$

Tomaremos 11 cabezas de ganado en promedio por cada explotación ganadera.

$$\hat{V}(\hat{R}) = \frac{1}{n\bar{x}^2}(\hat{S}_y^2 + \hat{R}^2\hat{S}_x^2 - 2\hat{R}\hat{S}_{xy}) = \frac{1}{10(4{,}4)^2}\left(795{,}51 + 11^2 \cdot 6{,}26 - 2 \cdot 11 \cdot 70{,}2\right) = 0{,}004426$$

$$\hat{Cv}(\hat{R}) = \frac{\hat{\sigma}(\hat{R})}{\hat{R}} = \frac{\sqrt{0{,}004426}}{11} = 0{,}006 \quad (0{,}6\%)$$

El muestral para afijación proporcional con reposición para un error relativo del 5% al estimar el total de cabezas de ganado se halla despejando n en la expresión:

$$0{,}1 = \hat{Cv}(\hat{Y}) = \frac{\hat{\sigma}(\hat{Y})}{\hat{Y}} = \frac{\sqrt{\sum_{h=1}^{2} \frac{N_h^2 \hat{S}_{yh}^2}{\frac{n}{N} N_h}}}{2780} = \frac{\sqrt{\frac{N}{n}\sum_{h=1}^{2} N_h \hat{S}_{yh}^2}}{2780} = \frac{\sqrt{\frac{253}{n}(71 \cdot 7{,}61 + 182 \cdot 30{,}15)}}{2780} \Rightarrow n \cong 20$$

La afijación será $n_1 = (20/253)71 = 6$ y $n_2 = (20/253)182 = 14$ (6 municipios del estrato 1 y 14 municipios el estrato 2).

6.8. Se trata de estudiar el ganado ovino en una determinada comarca en la que existen seis majadas. Para ello se estratifica la comarca en dos zonas, una de secano a la que corresponden tres majadas y otra de regadío a la que corresponden las otras tres majadas. En cada majada se mide el número de ovejas (variable X) y su superficie en unidades cuadradas (variable Y), y se obtienen los siguientes datos:

Estrato 1		*Estrato* 2	
X_{1i}	Y_{1i}	X_{2i}	Y_{2i}
2	1	5	4
4	2	7	5
5	3	12	6

A la vista de la información, analizar la precisión de todos los métodos indirectos de estimación que se utilizan en estratificación cuando se trata de estimar el número medio de ovejas por majada utilizando la información adicional de la variable auxiliar Y. Razonar adecuadamente los resultados. Contrastar también estos resultados con las precisiones obtenidas considerando métodos de estimación indirecta sin estratificación. Emplear también métodos directos de estimación para la variable en estudio sin utilizar la variable de apoyo.

Los métodos de estimación indirecta son perfectamente aplicables en este caso porque el coeficiente de correlación entre la variable en estudio X y la variable auxiliar Y es muy alto (0,9).

A partir de los datos del problema se puede construir la siguiente tabla:

Estrato	N_h	W_h	S_{xh}^2	S_{yh}^2	$\overline{X}_h$	$\overline{Y}_h$	S_{xyh}	f_h	n_h
1	3	1/2	7/3	1	11/3	2	3/2	2/3	2
2	3	1/2	13	1	8	5	7/2	2/3	2

A continuación se calculan las varianzas del estimador de la media para los distintos métodos de estimación directos e indirectos y estratificados y sin estratificar.

$$\textit{Aleatorio simple} \rightarrow V_1(\overline{x}) = (1-f)\frac{S_x^2}{n} = 0{,}98$$

$$\textit{Estratificado} \rightarrow V_2(\overline{x}) = \sum_{h=1}^{2} W_h^2 (1-f_h)\frac{S_{xh}^2}{n_h} = 0{,}63$$

$$\textit{Razón} \rightarrow V_3(\overline{x}) = \frac{(1-f)}{n}\left(S_x^2 + R^2 S_y^2 - 2RS_{xy}\right) = 0{,}151296$$

$$\textit{Razón separada} \rightarrow V_4(\overline{x}) = \sum_{h=1}^{2} W_h^2 \frac{(1-f_h)}{n}\left(S_{xh}^2 + R_h^2 S_{yh}^2 - 2R_h S_{xyh}\right) = 0{,}189$$

$$\textit{Razón combinada} \rightarrow V_5(\overline{x}) = \sum_{h=1}^{2} W_h^2 \frac{(1-f_h)}{n}\left(S_{xh}^2 + R^2 S_{yh}^2 - 2RS_{xyh}\right) = 0{,}1759$$

$$\textit{Regresión} \rightarrow V_6(\overline{x}) = (1-f)\frac{S_x^2}{n}(1-\rho^2) = 0{,}15119$$

$$\textit{Regresión separada} \rightarrow V_7(\bar{x}) = \sum_{h=1}^{2} W_h^2 \frac{(1-f_h)}{n}\left(S_{xh}^2 + \beta_h^2 S_{yh}^2 - 2\beta_h S_{xyh}\right) = 0{,}0347$$

$$\textit{Regresión combinada} \rightarrow V_8(\bar{x}) = \sum_{h=1}^{2} W_h^2 \frac{(1-f_h)}{n}\left(S_{xh}^2 + \overline{\beta}_c^2 S_{yh}^2 - 2\overline{\beta}_c S_{xyh}\right) = 0{,}118$$

$$\textit{Diferencia} \rightarrow V_9(\bar{x}) = \frac{(1-f)}{n}\left(S_x^2 + S_y^2 - 2S_{xy}\right) = 0{,}28833$$

En cuanto a los métodos no estratificados, se observa que la estimación óptima la produce el método indirecto basado en la regresión, resultado que siempre se cumple.

El siguiente método en precisión es la estimación indirecta por razón, que presenta una precisión muy similar a la estimación por regresión (apenas un 0,07% de ganancia en precisión para regresión).

La estimación indirecta por diferencia también es aceptable, aunque es el método de estimación indirecta menos preciso en este caso. Por otra parte, el muestreo aleatorio simple presenta una precisión muy inferior a cualquier método indirecto.

Ello nos lleva a concluir que en este problema es importante la consideración de los métodos indirectos de estimación.

Si analizamos la recta de regresión de la variable en estudio X respecto de la variable auxiliar Y, que tiene de ecuación $x = 1{,}6y - 0{,}06$, vemos que prácticamente pasa por el origen, razón por la cual el estimador por razón es muy preciso.

Además, la pendiente de la recta no está lejos de la unidad, con lo que la estimación indirecta por diferencia puede resultar también apropiada.

Por otra parte se cumple:

$$0{,}9 = \rho > \frac{1}{2} R \frac{S_y}{S_x} = 0{,}45$$

lo que indica que el muestreo aleatorio simple va a ser bastante menos preciso que el método de estimación por razón.

Al introducir la estratificación se obtiene buena mejora en la estimación indirecta por regresión separada y no tanto en la combinada (que ya sabemos que siempre es peor que la separada).

En cuanto a la estratificación por razón, se obtienen peores precisiones que cuando se usa razón sin estratificar. Por lo tanto, la estimación estratificada basada en la razón no es conveniente. De todas formas, la estimación por razón combinada resulta aquí más precisa que la estimación por razón separada.

6.9. Antes del ingreso en un centro educativo se hizo un examen de conocimientos matemáticos a 486 estudiantes. Se seleccionó una muestra irrestricta aleatoria de $n = 10$ estudiantes y se observaron sus progresos en cálculo mediante una prueba de conocimientos cuyas calificaciones constituyen la variable *Y*. Más adelante se observaron sus calificaciones finales en cálculo mediante la variable *X*. Los datos se recogen en la tabla siguiente:

Estudiante	x	y
1	39	65
2	43	78
3	21	52
4	64	82
5	57	92
6	47	89
7	28	73
8	75	98
9	34	56
10	52	75

Se sabe que la calificación media de la prueba de conocimientos para los 486 estudiantes que presentaron el examen es 52. Estimar la calificación final media en cálculo para esta población, y establecer un límite para el error de estimación.

A fin de aprovechar la información adicional de la variable *Y*, para estimar la media de *X* utilizaremos el método de estimación indirecta más preciso, que es el estimador por regresión. Podemos resumir las estimaciones por regresión como sigue:

$$\bar{x}_{rg} = \bar{x} + b_o\left(\bar{Y} - \bar{y}\right)$$

Del enunciado del problema sabemos que $\bar{Y} = 52$, y de los datos de la tabla se deduce que $\bar{x} = 76$ e $\bar{y} = 46$. Para calcular el estimador por regresión sólo nos faltaría estimar b_o. Tenemos:

$$\hat{b}_0 = \hat{\beta} = \frac{\hat{S}_{XY}}{\hat{S}_Y^2} = \frac{\sum_i^n \left(X_i - \bar{x}\right)\left(Y_i - \bar{y}\right)}{\sum_i^n \left(Y_i - \bar{y}\right)^2} = \frac{\sum_i^n X_i Y_i - n\bar{x}\bar{y}}{\sum_i^n Y_i^2 - n\bar{y}^2} = \frac{36{,}854 - 10(46)(76)}{23{,}634 - 10(46)^2} = 0{,}766$$

El estimador por regresión será entonces:

$$\bar{x}_{rg} = \bar{x} + b_o\left(\bar{Y} - \bar{y}\right) = 76 + 0{,}766(52 - 46) = 80$$

La varianzas mínima estimada será $\hat{V}_{\min}(\bar{x}_{rg}) = \frac{(1-f)}{n} \cdot \hat{S}_x^2\left(1 - \hat{\rho}^2\right) = 7{,}4$ y el límite para el error de estimación al 95% es $2\sqrt{\hat{V}_{\min}(\bar{x}_{rg})} = 5{,}4$.

6.10. Los auditores frecuentemente están interesados en comparar el valor intervenido de los artículos con el valor asentado en los libros. Generalmente, los valores en los libros son conocidos para cada artículo en la población, y los valores intervenidos son obtenidos con una muestra de esos artículos. Los valores en el libro entonces pueden utilizarse para obtener una buena estimación del valor intervenido total o promedio para la población. Supóngase que una población contiene 180 artículos inventariados con un valor establecido en el libro de \$13,320. Denotar por y_i el valor en el libro y por x_i el valor intervenido del i-ésimo artículo. Una muestra irrestricta aleatoria de $n = 10$ artículos produce los resultados que se muestran en la tabla adjunta. Estimar el valor intervenido medio por el método de diferencia así como el error cometido. Realizar las mismas estimaciones pero usando un estimador de regresión y un estimador de razón.

Muestra	Valor intervenido	Valor en el libro	d_i
	x_i	y_i	
1	9	10	-1
2	14	12	2
3	7	8	-1
4	29	26	3
5	45	47	-2
6	109	112	-3
7	40	36	4
8	238	240	-2
9	60	59	1
10	170	167	3

La estimación por diferencia se realiza de la siguiente forma:

$$\hat{\bar{X}} = \bar{x} - \bar{y} + \bar{Y} = \hat{D} + \bar{Y} = (72{,}1\text{-}71{,}7) + 74 = 74{,}4$$

La estimación de la varianza viene dada por:

$$\hat{V}(\hat{\bar{X}}) = \frac{1-f}{n}\left(\hat{S}_x^2 + \hat{S}_y^2 - 2\hat{S}_{xy}\right) = 0{,}59$$

La estimación por regresión se realiza de la siguiente forma:

$$\bar{x}_{rg} = \bar{x} + b_o\left(\bar{Y} - \bar{y}\right) = 72{,}1 + 0{,}99(74 - 71{,}7) = 74{,}38$$

$$\hat{b}_0 = \frac{\sum_i^n \left(X_i - \bar{x}\right)\left(Y_i - \bar{y}\right)}{\sum_i^n \left(Y_i - \bar{y}\right)^2} = \frac{\sum_i^n X_i Y_i - n\bar{x}\bar{y}}{\sum_i^n Y_i^2 - n\bar{y}^2} = \frac{105{,}881 - 10(71{,}7)(72{,}1)}{106{,}003 - 10(71{,}7)^2} = 0{,}99$$

La varianzas mínima estimada será $\hat{V}_{\min}(\bar{x}_{rg}) = \frac{(1-f)}{n} \cdot \hat{S}_x^2 \left(1 - \hat{\rho}^2\right) = 2{,}24$.

La estimación por regresión se realiza de la siguiente forma:

$$\hat{\bar{X}}_R = \bar{x}_R = \frac{\bar{x}}{\bar{y}} \bar{Y} = \hat{R}\bar{Y} = \frac{721}{717} 74 = 74{,}41$$

La varianza puede estimarse como sigue:

$$\hat{V}(\hat{\bar{X}}_R) = \frac{1-f}{n}\left(\hat{S}_x^2 + \hat{R}^2 \hat{S}_y^2 - 2\hat{R}\hat{S}_{xy}\right) = \frac{1-f}{n(n-1)}\left[\sum_i^n X_i^2 + \hat{R}^2 \sum_i^n Y_i^2 - 2\hat{R}\sum_i^n X_i Y_i\right] = 0{,}66$$

EJERCICIOS PROPUESTOS

6.1. Sobre una población de 500 unidades está definida un característica bidimensional (X_i, Y_i). Una muestra aleatoria simple de tamaño 80 proporciona los siguientes datos:

$$\sum_{i=1}^{80} X_i = 420 \quad , \quad \sum_{i=1}^{80} Y_i = 190 \quad , \quad \sum_{i=1}^{80} X_i^2 = 2284 \quad , \quad \sum_{i=1}^{80} Y_i^2 = 512 \quad \text{y} \quad \sum_{i=1}^{80} X_i Y_i = 1045$$

a) Estimar el sesgo y el error de muestreo de la razón de la variable *Y* a la variable *X*. ¿Se trata de un sesgo influyente para estimaciones indirectas basadas en la razón?

b) Se trata de estimar con y sin reposición la media y el total de *Y* utilizando la información adicional de la variable *X* mediante un método de estimación indirecta. ¿Qué método indirecto sería el más adecuado? ¿Por qué? Realizar las estimaciones de media y total mediante los métodos indirectos conocidos ordenándolos en precisión y sabiendo que el total de *X* es 10000.

c) ¿Habrá ganancia en precisión respecto del muestreo aleatorio simple? Cuantificarla.

6.2. Una empresa está interesada en estimar el total de ganancias por las ventas de televisiones de color al final de un período de tres meses (variable *Y*). Se tienen cifras del total de ganancias de todas las sucursales de la empresa para el período de tres meses correspondiente del año anterior (variable *X*). Se selecciona una muestra irrestricta aleatoria de 13 sucursales de entre las 123 de la empresa. Usando un estimador de razón, estimar el total de ganancias por las ventas de televisiones de color al final de un período de tres meses y establecer un límite para el error de estimación. Usar los datos de la tabla adjunta, y considerar que la media poblacional de la variable *X* vale 128,200.

Oficina	Datos de tres meses del año anterior, *Xi*	Datos de tres meses del año actual, *Yi*
1	550	610
2	720	780
3	1500	1600
4	1020	1030
5	620	600
6	980	1050
7	928	977
8	1200	1440
9	1350	1570
10	1750	2210
11	670	980
12	729	865
13	1530	1710

Estimar también las ganancias medias para las oficinas de la empresa y establecer un límite para el error de estimación.

6.3. Una empresa industrial elabora un producto que es empaquetado, para propósitos de mercado, en dos marcas comerciales. Estas dos marcas sirven como estratos para estimar el volumen potencial de ventas para el trimestre siguiente. Una muestra irrestricta aleatoria de clientes para cada marca es entrevistada para proporcionar una cantidad potencial Y de ventas (en número de unidades) para el próximo trimestre. La cifra de las ventas verdaderas del año pasado, para el mismo trimestre, está disponible para cada uno de los clientes muestreados y se denota por X. Los datos se presentan en la tabla anexa. La muestra para la marca I fue tomada de una lista de 120 clientes, para quienes el total de ventas en el mismo trimestre del año pasado fue de 24500 unidades. La muestra de la marca II viene de 180 clientes, con un total trimestral de ventas para el año pasado de 21000 unidades. Hallar una estimación de razón del total potencial de ventas para el próximo trimestre. Estime la varianza de su estimador.

Marca I		Marca II	
Xi	*Yi*	*Xi*	*Yi*
204	210	137	150
143	160	189	200
82	75	119	125
256	280	63	60
275	300	103	110
198	190	107	100
		159	180
		63	75
		87	90

6.4. Se estima el ingreso nacional para 1981 mediante una muestra de $n = 10$ industrias que declaran sus ingresos de 1981 antes que las 35 restantes. Se dispone de los datos del ingreso de 1980 para las 45 industrias y los totales son 2174,2 (en miles de millones). Los datos se presentan en la tabla adjunta.

Industria	1980	1981
Productos de fábricas textiles	13,6	14,5
Productos químicos y relacionados	37,7	42,7
Madera aserrada y leña	15,2	15,1
Equipo eléctrico y electrónico	48,4	53,6
Vehículos automotores y equipo	19,6	25,4
Transporte y almacenaje	33,5	35,9
Banca	44,4	48,5
Bienes raíces	198,3	221,2
Servicios de salud	99,2	114
Servicios de educación	15,4	17

a) Hallar un estimador de razón del ingreso total de 1981, y establecer un límite para el error de estimación.
b) Hallar un estimador de regresión del ingreso total de 1981, y establecer un límite para el error de estimación.
c) hallar un estimador de diferencia del ingreso total de 1981, y establecer un límite para el error de estimación.
d) ¿Cuál de los tres métodos es el más apropiado en este caso? ¿Por qué?

CAPÍTULO

7

MUESTREO UNIETÁPICO DE CONGLOMERADOS

OBJETIVOS

1. Presentar el concepto de muestreo unietápico de conglomerados.
2. Analizar los estimadores y sus errores en muestreo unietápico de conglomerados del mismo tamaño y con probabilidades iguales.
3. Analizar los errores y su estimación en función del coeficiente de correlación intraconglomerados.
4. Analizar los estimadores y sus errores cuando se considera muestreo unietápico de conglomerados con reposición.
5. Estudiar el muestreo unietápico de conglomerados de distinto tamaño y probabilidades iguales con y sin reposición.
6. Estudiar el muestreo unietápico de conglomerados de distinto tamaño y probabilidades desiguales con y sin reposición.
7. Estudiar el muestreo unietápico de conglomerados de distinto tamaño y probabilidades proporcionales al tamaño con y sin reposición.
8. Estudiar el problema del tamaño de la muestra.

ÍNDICE

1. Muestreo unietápico de conglomerados. Estimadores para conglomerados del mismo tamaño y probabilidades iguales.

2. Varianza de los estimadores. Coeficiente de correlación intraconglomerados. Estimación de varianzas.

3. Muestreo de conglomerados del mismo tamaño con reposición. Varianzas de los estimadores y estimación de las varianzas.

4. Muestreo unietápico de conglomerados de distinto tamaño.

5. Muestreo unietápico de conglomerados de distinto tamaño con probabilidades desiguales.

6. Tamaño de la muestra.

7. Problemas resueltos.

8. Ejercicios propuestos.

MUESTREO UNIETÁPICO DE CONGLOMERADOS. ESTIMADORES PARA CONGLOMERADOS DEL MISMO TAMAÑO Y PROBABILIDADES IGUALES

Tanto en el muestreo aleatorio simple con reposición como sin reposición, así como en el muestreo estratificado, sistemático y métodos indirectos de estimación, las unidades de muestreo son las mismas que las unidades objeto de estudio (unidades simples o elementales), pero en la práctica nos encontramos con situaciones más generales en las que las unidades de muestreo comprenden dos o más unidades de estudio. En tal caso a las unidades de muestreo se las denomina unidades primarias o compuestas.

En el muestreo por conglomerados no se necesita un marco muy específico como en el caso del muestreo aleatorio simple en el que era necesario disponer de un listado de unidades de la población, o como en el muestreo estratificado, donde era necesario disponer de listados de unidades por estratos. Se divide previamente al muestreo la población en conglomerados o áreas convenientes, de las cuales se selecciona un cierto número para la muestra, con lo que sólo es necesario un marco de conglomerados que será más fácil de conseguir y más barato. Se pueden utilizar como marco divisiones territoriales ya establecidas por necesidades administrativas para las cuales existe ya información. También se pueden utilizar como marco áreas geográficas cuyas características están ya muy delimitadas. Está claro que se ahorra coste y tiempo al efectuar visitas a las unidades seleccionadas. Además, la concentración de unidades disminuye la necesidad de desplazamientos.

Por otro lado, en el muestreo por conglomerados solemos tener menor precisión en las estimaciones, debido a que, aunque lo ideal es que haya heterogeneidad dentro, siempre va a existir un cierto grado de homogeneidad inevitable dentro de los conglomerados que disminuirá la precisión. La eficiencia de este tipo de muestreo disminuye al aumentar el tamaño de los conglomerados, cuando en realidad este tipo de muestreo es más útil en caso de poblaciones muy numerosas en las que se puedan construir conglomerados grandes.

Consideramos una población finita con M unidades elementales o últimas agrupadas en N unidades mayores llamadas conglomerados o unidades primarias, de tal forma que no existan solapamientos entre los conglomerados y que éstos contengan en todo caso a la población en estudio. Consideramos como unidad de muestreo el conglomerado, y extraemos de la población una muestra de n conglomerados a partir de la cual estimaremos los parámetros poblacionales. El número de unidades elementales de un conglomerado se denomina tamaño del conglomerado. Los conglomerados pueden ser de igual o de distinto tamaño, y han de ser lo más heterogéneos posible dentro de ellos y lo más homogéneos posible entre ellos, de tal forma que la situación ideal sería que un único conglomerado pudiese representar fielmente a la población (muestra de tamaño uno con mínimo coste). Se observa que la situación ahora es la complementaria a la del caso de los estratos estudiados anteriormente.

Vamos a suponer ahora *probabilidades iguales y que todos los conglomerados son del mismo tamaño* $\overline{M}$, en cuyo caso utilizaremos la siguiente notación:

N: Número de conglomerados en la población
N: Número de conglomerados en la muestra
$\overline{M}$: Número de unidades elementales por conglomerado (tamaño del conglomerado)
$N\overline{M}$: Número total de unidades elementales en la población
$n\overline{M}$: Número total de unidades elementales en la muestra

Consideraremos la característica poblacional general $\theta = \sum_i^N Y_i = \sum_i^N \sum_i^{\overline{M}} Y_{ij}$ que, suponiendo *muestreo sin reposición y probabilidades iguales*, puede ser estimada mediante el estimador lineal insesgado de Horwitz y Thompson $\hat{\theta}_{HT} = \sum_i^n \frac{Y_i}{\pi_i} = \sum_i^n \frac{\sum_i^{\overline{M}} Y_{ij}}{n/N} = \frac{N}{n} \sum_i^n \sum_i^{\overline{M}} Y_{ij}$.

La aplicación del estimador lineal insesgado de Horwitz y Thompson para probabilidades iguales a las estimaciones del total, media, proporción y total de clase poblacionales, proporciona los siguientes estimadores:

$$\theta = X = \sum_i^N \sum_j^{\overline{M}} X_{ij} \Rightarrow Y_{ij} = X_{ij} \Rightarrow \hat{X} = \frac{N}{n} \sum_i^n \sum_j^{\overline{M}} X_{ij} = \frac{N\overline{M}}{n} \sum_i^n \frac{1}{\overline{M}} \sum_j^{\overline{M}} X_{ij} = N\overline{M} \frac{1}{n} \sum_i^n \overline{X}_i = N\overline{M}\overline{\overline{x}}$$

$$\theta = \overline{X} = \frac{1}{N\overline{M}} \sum_i^N \sum_j^{\overline{M}} X_{ij} \Rightarrow Y_{ij} = \frac{X_{ij}}{N\overline{M}} \Rightarrow \hat{\overline{X}} = \frac{N}{n} \sum_i^n \sum_j^{\overline{M}} \frac{X_{ij}}{N\overline{M}} = \frac{1}{n} \sum_i^n \frac{1}{\overline{M}} \sum_j^{\overline{M}} X_{ij} = \frac{1}{n} \sum_i^n \overline{X}_i = \overline{\overline{x}}$$

$$\theta = P = \frac{1}{N\overline{M}} \sum_i^N \sum_j^{\overline{M}} A_{ij} \Rightarrow Y_{ij} = \frac{A_{ij}}{N\overline{M}} \Rightarrow \hat{P} = \frac{N}{n} \sum_i^n \sum_j^{\overline{M}} \frac{A_{ij}}{N\overline{M}} = \frac{1}{n} \sum_i^n \frac{1}{\overline{M}} \sum_j^{\overline{M}} A_{ij} = \frac{1}{n} \sum_i^n P_i$$

$$\theta = A = \sum_i^N \sum_j^{\overline{M}} A_{ij} \Rightarrow Y_{ij} = A_{ij} \Rightarrow \hat{A} = \frac{N}{n} \sum_i^n \sum_j^{\overline{M}} A_{ij} = \frac{N\overline{M}}{n} \sum_i^n \frac{1}{\overline{M}} \sum_j^{\overline{M}} A_{ij} = N\overline{M} \frac{1}{n} \sum_i^n P_i = N\overline{M}\hat{P}$$

VARIANZAS DE LOS ESTIMADORES. COEFICIENTE DE CORRELACIÓN INTRACONGLOMERADOS. ESTIMACIÓN DE LAS VARIANZAS

Las expresiones iniciales para las *varianzas de los estimadores sin reposición y probabilidades iguales para conglomerados del mismo tamaño* son:

$$V\left(\overline{\overline{x}}\right) = (1-f) \cdot \frac{S_b^2}{n\overline{M}} \quad \text{con } S_b^2 = \frac{\sum_i^N \sum_j^{\overline{M}} \left(\overline{X}_i - \overline{X}\right)^2}{N-1}$$

$$V\left(\hat{X}\right) = V\left(N\overline{M} \cdot \overline{\overline{x}}\right) = N^2 \overline{M}^2 \cdot V\left(\overline{\overline{x}}\right) = N^2 \overline{M}^2 \cdot (1-f) \cdot \frac{S_b^2}{n\overline{M}}$$

$$V\left(\hat{P}\right) = \left(1-f\right) \cdot \frac{\frac{\overline{M}}{N-1} \sum_i^N \left(P_i - P\right)^2}{n\overline{M}} = (1-f) \frac{\sum_i^N \left(P_i - P\right)^2}{n(N-1)}$$

$$V(\hat{A}) = V(N\overline{M} \cdot \hat{P}) = N^2 \overline{M}^2 V\left(\hat{P}\right) = N^2 \overline{M}^2 \cdot (1-f) \frac{\sum_i^N \left(P_i - P\right)^2}{n(N-1)}$$

Las expresiones de las varianzas son similares a las obtenidas en el muestreo aleatorio simple, sustituyendo S^2 por S_b^2 y siendo $n\overline{M}$ el número total de unidades elementales en la muestra.

Pero las varianzas anteriores pueden expresarse en función del *coeficiente de correlación intraconglomerados*, que se define como el coeficiente de correlación lineal entre todos los pares de valores de la variable en estudio medidos sobre las unidades de los conglomerados y extendido a todos los conglomerados, de tal forma que dicho coeficiente será una <<medida de la homogeneidad>> en el interior de los conglomerados. Evidentemente interesará que el coeficiente de homogeneidad intraconglomerados sea lo más pequeño posible, ya que en muestreo por conglomerados lo ideal es la heterogeneidad dentro de los conglomerados. La expresión del coeficiente de correlación intraconglomerados será:

$$\delta = \frac{Cov(X_{ij}, X_{iz})}{\sigma(X_{ij})\sigma(X_{iz})} = \frac{E\left[(X_{ij} - E(X_{ij}))(X_{iz} - E(X_{iz}))\right]}{\sigma^2} = \frac{\dfrac{1}{N\binom{\overline{M}}{2}}\sum_{i=1}^{N}\sum_{j<z}^{\overline{M}}(X_{ij} - \overline{X})(X_{iz} - \overline{X})}{\sigma^2}$$

de donde al ser $S^2 = \dfrac{1}{N\overline{M} - 1}\sum_{i}^{N}\sum_{j\neq l}^{\overline{M}}\left(X_{ij} - \overline{X}\right)^2$ y $\sigma^2 = \dfrac{1}{N\overline{M}}\sum_{i}^{N}\sum_{j\neq l}^{\overline{M}}\left(X_{ij} - \overline{X}\right)^2$ se puede expresar la varianza como $\sigma^2 = \dfrac{N\cdot\overline{M} - 1}{N\cdot\overline{M}}S^2$, expresión que puede sustituirse en el denominador del coeficiente de correlación intraconglomerados:

$$\delta = \frac{\dfrac{1}{N\binom{\overline{M}}{2}}\sum_{i=1}^{N}\sum_{j<z}^{\overline{M}}(X_{ij} - \overline{X})(X_{iz} - \overline{X})}{\dfrac{N\cdot\overline{M} - 1}{N\cdot\overline{M}}S^2} = \frac{2\sum_{i=1}^{N}\sum_{j<z}^{\overline{M}}(X_{ij} - \overline{X})(X_{iz} - \overline{X})}{\left(\overline{M} - 1\right)\left(N\overline{M} - 1\right)S^2}$$

Este coeficiente se puede estimar mediante $\hat{\delta} = \dfrac{\hat{S}_b^2 - \hat{S}^2}{\left(\overline{M} - 1\right)\hat{S}_0{}^2}$

$$\hat{S}^2 = \frac{1}{n\overline{M} - 1}\sum_{i}^{n}\sum_{j\neq l}^{\overline{M}}\left(X_{ij} - \overline{\overline{x}}\right)^2,\quad \hat{S}_w^2 = \frac{1}{n\overline{M} - n}\sum_{i}^{n}\sum_{j}^{\overline{M}}\left(X_{ij} - \overline{X}_i\right)^2,\quad \hat{S}_b^2 = \frac{1}{n-1}\sum_{i}^{n}\sum_{j}^{\overline{M}}\left(\overline{X}_i - \overline{\overline{x}}\right)^2$$

$$\hat{S}_0{}^2 = \frac{N-1}{N\overline{M} - 1}\cdot\hat{S}_b^2 + \frac{N\left(\overline{M} - 1\right)}{N\overline{M} - 1}\cdot\hat{S}_w^2 \qquad \hat{S}^2 = \frac{n-1}{n\overline{M} - 1}\cdot\hat{S}_b^2 + \frac{n\left(\overline{M} - 1\right)}{n\overline{M} - 1}\cdot\hat{S}_w^2$$

Los errores de estos estimadores y sus estimaciones en función de ρ son:

$$V\left(\overline{\overline{x}}\right) = \left(1 - f\right)\frac{S^2}{n\overline{M}}\left[1 + \left(\overline{M} - 1\right)\delta\right] \Rightarrow \hat{V}\left(\overline{\overline{x}}\right) = \left(1 - f\right)\frac{\hat{S}_0{}^2}{n\overline{M}}\left[1 + \left(\overline{M} - 1\right)\hat{\delta}\right]$$

$$V\left(\overline{\overline{x}}\right) = \left(1 - f\right)\frac{S_b^2}{n\overline{M}} \Rightarrow \hat{V}\left(\overline{\overline{x}}\right) = \left(1 - f\right)\frac{\hat{S}_b^2}{n\overline{M}}$$

$$V(\hat{X}) = V(N\overline{M}\overline{x}) = N^2\overline{M}^2V(\overline{\overline{x}}) \Rightarrow \hat{V}(\hat{X}) = N^2\overline{M}^2\hat{V}(\overline{\overline{x}})$$

El cálculo de los términos de las fórmulas anteriores los facilitan los cuadros del análisis de la varianza para la población y para la muestra siguientes:

Descomposición de la varianza para la población

Fuente de variación	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*
Entre conglomerados	$N-1$	$\sum_{i}^{N}\sum_{j}^{\overline{M}}\left(\overline{X}_i-\overline{\overline{X}}\right)^2$	S_b^2
Dentro de conglomerados	$n\left(\overline{M}-1\right)$	$\sum_{i}^{N}\sum_{j}^{\overline{M}}\left(X_{ij}-\overline{X}_i\right)^2$	S_w^2
Total	$N\overline{M}-1$	$\sum_{i}^{N}\sum_{j}^{\overline{M}}\left(X_{ij}-\overline{\overline{X}}\right)^2$	

Descomposición de la varianza para la muestra

Fuente de variación	*Grados de libertad*	*Sumas de cuadrados*	*Cuadrados medios*	*Esperanzas*
Entre conglomerados	$n-1$	$\sum_{i}^{n}\sum_{j}^{\overline{M}}\left(\overline{X}_i-\overline{\overline{x}}\right)^2$	$\hat{S}_b^2$	S_b^2
Dentro de conglom.	$n\left(\overline{M}-1\right)$	$\sum_{i}^{n}\sum_{j}^{\overline{M}}\left(X_{ij}-\overline{X}_i\right)^2$	$\hat{S}_w^2$	S_w^2
Total	$n\overline{M}-1$	$\sum_{i}^{n}\sum_{j}^{\overline{M}}\left(X_{ij}-\overline{\overline{x}}\right)^2$	$\hat{S}^2$	

Para el caso de *proporciones y totales de clase* las fórmulas son las mismas, pero las magnitudes se obtienen del cuadro del análisis de la varianza siguiente:

Fuentede Variación	*Gradosde Libertad*	*Sumade Cuadrados*	*Cuadrados Medios*	*Estimadores Insesgados*
Entre	$N-1$	$A=\sum_{i=1}^{N}\overline{M}(P_i-P)^2$	$S_b^2=\frac{A}{N-1}$	$\hat{S}_b^2=\frac{\sum_{i=1}^{N}\overline{M}(P_i-\frac{1}{n}\sum_{i=1}^{n}P_i)^2}{n-1}$
Dentro	$N(\overline{M}-1)$	$B=\sum_{i=1}^{N}\overline{M}P_i(1-P_i)$	$S_w^2=\frac{B}{N(\overline{M}-1)}$	$\hat{S}_w^2=\frac{\sum_{i=1}^{n}\overline{M}P_i(1-P_i)}{n(\overline{M}-1)}$
Total	$N\overline{M}-1$	$C=N\overline{M}P(1-P)$	$S^2=\frac{C}{N\overline{M}-1}$	$\hat{S}_0^2$

Comparación con el muestreo aleatorio simple

De la expresión $V(\overline{\overline{x}})=(1-f)\frac{S^2}{n\overline{M}}\left[1+(\overline{M}-1)\cdot\delta\right]=V_{MAS}(\overline{x})\left[1+(\overline{M}-1)\cdot\delta\right]$ se deduce que para valores positivos de δ existe un aumento en la varianza del muestreo por conglomerados con relación al muestreo aleatorio simple y muestras de tamaño igual a $n\cdot\overline{M}$ unidades elementales.

El caso más desfavorable (varianza máxima) correspondería a $\delta = +1$ y el más favorable (varianza mínima) a $\delta = -\frac{1}{\overline{M}-1}$, en que la varianza sería igual a cero. Para $\delta = 0$ ambos métodos proporcionarían la misma precisión.

El término $\overline{M}-1$ expresa el aumento de la varianza debido a la selección de n conglomerados de tamaño $\overline{M}$ en lugar de $n\,\overline{M}$ unidades elementales obtenidas por muestreo aleatorio simple. Ahora bien, si el coeficiente de correlación intraconglomerados fuese negativo, ello supondría mayor precisión en el muestreo por conglomerados que en el aleatorio simple.

Pero en la práctica suele ocurrir que los elementos de cada conglomerado tienen cierto parecido entre sí aunque se intente que sean lo más heterogéneos posible, con lo cual la correlación es positiva y menor la precisión en el muestreo por conglomerados que en el aleatorio simple. Este problema ya se había citado al principio del capítulo como una de las desventajas del muestreo por conglomerados.

Según lo visto, la comparación entre muestreo monoetápico de conglomerados y muestreo aleatorio simple podría resumirse como sigue:

$$V_{MC}\left(\bar{\bar{x}}\right) = V_{MAS}\left(\bar{x}\right)\cdot\left[1+\left(\overline{M}-1\right)\cdot\delta\right] \Rightarrow \begin{cases} Si\,\delta > 0 \Rightarrow \textit{conglomerados peor que aleatorio simple} \\ Si\,\delta = 0 \Rightarrow \textit{conglomerados igual que aleatorio simple} \\ Si\,\delta < 0 \Rightarrow \textit{conglomerados mejor que aleatorio simple} \end{cases}$$

Evidentemente, cuando $\delta \in (0,1]$ la precisión del muestreo por conglomerados es inferior a la del muestreo aleatorio simple, y a medida que el δ se aproxima a 1, se acentúa la pérdida de precisión en el muestreo por conglomerados respecto del aleatorio simple. Cuando $\delta = 0$, las precisiones de ambos métodos coinciden, y cuando $\delta \in \left[-\frac{1}{\overline{M}-1}, 0\right]$, la precisión del muestreo por conglomerados es superior a la del muestreo aleatorio simple y a medida que el δ se aproxima a $-\frac{1}{\overline{M}-1}$, se acentúa la ganancia en precisión del muestreo por conglomerados respecto del aleatorio simple.

Por otra parte, si llamamos n_a al tamaño de muestra necesario en muestreo aleatorio simple para obtener una precisión dada, y si llamamos n_c al tamaño de muestra en muestreo por conglomerados, resulta que si los dos tipos de muestreo tienen la misma precisión, $(1-f)\frac{S^2}{n_a} = (1-f)\frac{S^2}{n_c}(1+(\overline{M}-1)\delta) \Rightarrow n_c = n_a(1+(\overline{M}-1)\delta)$.

Precisamente la cantidad $1+\left(\overline{M}-1\right)\cdot\delta$ por la que hay que multiplicar el tamaño de una muestra por conglomerados n_c para que coincida con el tamaño de muestra necesario en muestreo aleatorio simple n_a para igual precisión en ambos tipos de muestreo, se denomina ***efecto del diseño***.

MUESTREO DE CONGLOMERADOS DEL MISMO TAMAÑO CON REPOSICIÓN. VARIANZAS DE LOS ESTIMADORES Y ESTIMACIÓN DE LAS VARIANZAS

En caso de *muestro con reposición, probabilidades iguales y conglomerados del mismo tamaño*, los estimadores son los mismos, y las varianzas tienen las siguientes expresiones:

$$V\left(\bar{\bar{x}}\right)=\frac{\frac{1}{N}\sum_{i}^{N}\overline{M}\left(\bar{X}_i-\bar{\bar{X}}\right)^2}{n\overline{M}}=\frac{\sigma_b^2}{n\overline{M}}$$

$\sigma_b^2=\frac{1}{N}\sum_{i}^{N}\overline{M}\left(\bar{X}_i-\bar{\bar{X}}\right)^2=\frac{1}{N}\sum_{i}^{N}\sum_{j}^{\overline{M}}\left(\bar{X}_i-\bar{\bar{X}}\right)^2$ es la cuasivarianza entre conglomerados y la expresión de la varianza de la media $V\left(\bar{\bar{x}}\right)=\frac{\sigma_b^2}{n\overline{M}}$ es similar a la obtenida en el muestreo aleatorio simple, sustituyendo σ^2 por σ_b^2 y siendo $n\overline{M}$ el número total de unidades elementales en la muestra.

$$V\left(\hat{X}\right)=V\left(N\overline{M}\cdot\bar{\bar{x}}\right)=N^2\overline{M}^2\cdot V\left(\bar{\bar{x}}\right)=N^2\overline{M}^2\frac{\sigma_b^2}{n\overline{M}}$$

$$V\left(\hat{P}\right)=\frac{\sigma_b^2}{n\overline{M}}=\frac{\frac{\overline{M}}{N}\sum_{i}^{N}\left(P_i-P\right)^2}{n\overline{M}}=\frac{\sum_{i}^{N}\left(P_i-P\right)^2}{nN}$$

$$V(\hat{A})=V(N\overline{M}\cdot\hat{P})=N^2\overline{M}^2V\left(\hat{P}\right)=N^2\overline{M}^2\frac{\sum_{i}^{N}\left(P_i-P\right)^2}{nN}$$

La varianzas de los estimadores y sus estimaciones en función del coeficiente de correlación intraconglomerados tienen las siguientes expresiones:

$$V\left(\bar{\bar{x}}\right)=\frac{\sigma^2}{n\overline{M}}\left[1+\left(\overline{M}-1\right)\delta\right]\Rightarrow\hat{V}\left(\bar{\bar{x}}\right)=\frac{\hat{\sigma}'^2}{n\overline{M}}\left[1+\left(\overline{M}-1\right)\hat{\delta}\right],\quad V\left(\bar{\bar{x}}\right)=\frac{\sigma_b^2}{n\overline{M}}\Rightarrow\hat{V}\left(\bar{\bar{x}}\right)=\frac{\hat{S}_b^2}{n\overline{M}}$$

$$V(\hat{X})=V(N\overline{M}\,\bar{\bar{x}})=N^2\overline{M}^2V(\bar{\bar{x}})\Rightarrow\hat{V}(\hat{X})=N^2\overline{M}^2\hat{V}(\bar{\bar{x}})$$

El *coeficiente de correlación intraconglomerados y su estimación* son:

$$\delta=\frac{\sigma_b^2-\sigma^2}{\left(\overline{M}-1\right)\sigma^2}\qquad\hat{\delta}=\frac{\hat{S}_b^2-\left(\hat{S}_{1w}^2+\frac{\hat{S}_b^2}{\overline{M}}\right)}{\left(\overline{M}-1\right)\left(\hat{S}_{1w}^2+\frac{\hat{S}_b^2}{\overline{M}}\right)}=\frac{\hat{S}_b^2-\hat{\sigma}'^2}{\left(\overline{M}-1\right)\hat{\sigma}'^2}$$

$$\hat{\sigma}'^2=\hat{S}_{1,w}^2+\frac{\hat{S}_b^2}{\overline{M}},\ \hat{S}_{1,w}^2=\frac{1}{n\overline{M}}\sum_{i}^{n}\sum_{j}^{\overline{M}}\left(X_{ij}-\bar{X}_i\right)^2,\ \sigma_w^2=\frac{1}{N\overline{M}}\sum_{i}^{N}\sum_{j}^{\overline{M}}\left(X_{ij}-\bar{X}_i\right)^2,\ \hat{S}_b^2=\frac{1}{n-1}\sum_{i}^{n}\sum_{j}^{\overline{M}}\left(\bar{X}_i-\bar{\bar{x}}\right)^2$$

Si *estimamos proporciones y totales de clase* utilizaremos lo siguiente:

$$\sigma^2 = \frac{N\overline{M}-1}{N\overline{M}}S^2 = \frac{N\overline{M}-1}{N\overline{M}}\frac{N\overline{M}P(1-P)}{N\overline{M}-1} = \frac{N\overline{M}P(1-P)}{N\overline{M}} = P(1-P)$$

$$\sigma_w^2 = \frac{1}{N\overline{M}}\sum_i^N\sum_j^{\overline{M}}\left(X_{ij}-\overline{X}_i\right)^2 = \frac{1}{N\overline{M}}\sum_i^N\overline{M}\left(P_i-P\right)^2 = \frac{1}{N}\sum_i^N\left(P_i-P\right)^2$$

$$\sigma_b^2 = \frac{1}{N}\sum_i^N\sum_j^{\overline{M}}\left(\overline{X}_i-\overline{X}\right)^2 = \frac{\overline{M}}{N}\sum_i^N\left(\overline{X}_i-\overline{X}\right)^2 .$$

$$\hat{\sigma}_b^2 = \hat{S}_b^2 = \frac{\overline{M}}{n-1}\sum_i^n\left(P_i-\overline{P}\right)^2$$

$$\hat{\sigma}_w^2 = \hat{S}_{1,w}^2 = \frac{1}{n\overline{M}}\sum_i^n\sum_j^{\overline{M}}\left(X_{ij}-\overline{X}_i\right)^2 = \frac{1}{n\overline{M}}\sum_{i=1}^n\overline{M}P_i\left(1-P_i\right) = \frac{1}{n}\sum_{i=1}^n P_i\left(1-P_i\right)$$

$$\hat{\sigma}'^2 = \hat{S}_{1,w}^2 + \frac{\hat{S}_b^2}{\overline{M}} = \frac{1}{n}\sum_{i=1}^n P_i\left(1-P_i\right)+\frac{\overline{M}}{n-1}\sum_i^n\left(P_i-\overline{P}\right)^2$$

MUESTREO UNIETÁPICO DE CONGLOMERADOS DE DISTINTO TAMAÑO

Probabilidades iguales

a) Los conglomerados no varían mucho en tamaño (M_i similares)

Consideraramos $\overline{M} = \sum_{i=1}^N \frac{M_i}{M}$ como la media de los tamaños M_i de los conglomerados y utilizamos todas las fórmulas estudiadas hasta ahora, tanto para muestreo con reposición como para muestreo sin reposición. No obstante, suelen considerarse las siguientes expresiones alternativas para los estimadores:

Muestreo sin reposición

Para la media se tiene

$$\overline{\overline{x}} = \frac{1}{n}\sum_{i=1}^n\overline{X}_i = \frac{1}{n}\sum_{i=1}^n\frac{X_i}{\overline{M}} = \frac{1}{n\overline{M}}\sum_{i=1}^n X_i \ ,\quad V(\overline{\overline{x}}) = \frac{1-f}{n\overline{M}^2}\cdot\frac{\sum_{i=1}^N\left(X_i-\overline{X}\right)^2}{N-1} \ ,\quad \hat{V}(\overline{\overline{x}}) = \frac{1-f}{n\overline{M}^2}\frac{\sum_{i=1}^n\left(X_i-\overline{\overline{x}}\right)^2}{n-1}$$

Para el total se tiene el estimador $\hat{X} = N\overline{M}\overline{\overline{x}} = N\overline{M}\frac{1}{n\overline{M}}\sum_{i=1}^n X_i = \frac{N}{n}\sum_{i=1}^n X_i$, que no depende de $\overline{M}$.

Su varianza y estimación de varianza tampoco dependen de $\overline{M}$. Tenemos:

$$V(\hat{X}) = N^2 \frac{1-f}{n} \cdot \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N-1}, \quad \hat{V}(\hat{X}) = N^2 \frac{1-f}{n} \frac{\sum_{i=1}^{n} \left(X_i - \overline{\overline{x}}\right)^2}{n-1}$$

Muestreo con reposición

Para muestreo con reposición la varianza y estimación de varianza para el *estimador de la media* pueden calcularse como sigue:

$$V(\overline{\overline{x}}) = \frac{1}{n\overline{M}^2} \cdot \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N}, \quad \hat{V}(\overline{\overline{x}}) = \frac{1}{n\overline{M}^2} \frac{\sum_{i=1}^{n} \left(X_i - \overline{\overline{x}}\right)^2}{n-1}$$

La varianza y estimación de varianza para el *estimador del total* no dependerán de $\overline{M}$ y pueden calcularse como sigue:

$$V(\hat{X}) = \frac{N^2}{n} \cdot \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N}, \quad \hat{V}(\hat{X}) = \frac{N^2}{n} \cdot \frac{\sum_{i=1}^{n} \left(X_i - \overline{\overline{x}}\right)^2}{n-1}$$

En caso de estimación de totales y proporciones se utilizan las fórmulas ya vistas anteriormente para conglomerados del mismo tamaño tomando $\overline{M} = \sum_{i=1}^{N} \frac{M_i}{M}$, tanto para muestreo sin reposición como para muestreo con reposición.

b) Los conglomerados varían mucho en tamaño (M_i no similares y $M = \sum_{i=1}^{N} M_i$)

Si los tamaños de los conglomerados son significativamente distintos, un estimador sesgado de la media es el estimador de razón:

$$\hat{\overline{X}} = \overline{\overline{x}} = \hat{R} = \frac{\sum_{i}^{n} X_i}{\sum_{i}^{n} M_i}$$

Muestreo sin reposición

Por ser un estimador de la razón, su varianza aproximada es:

$$V(\overline{\overline{x}}) = (1-f) \cdot \frac{N^2}{nM^2} \frac{\sum_{i}^{N} M_i^2 \left(\overline{X}_i - \overline{X}\right)}{N-1}, \quad \hat{V}(\overline{\overline{x}}) = \hat{V}\left(\hat{R}\right) = (1-f) \cdot \frac{N^2}{nM^2} \frac{\sum_{i}^{n} M_i^2 \left(\overline{X}_i - \overline{\overline{x}}\right)^2}{n-1}$$

Para el estimador del total tendremos:

$$V\left(\hat{X}\right) = \frac{N^2(1-f)}{n} \frac{\sum_{i}^{N} M_i^2 \left(\overline{X}_i - \overline{X}\right)^2}{N-1}, \quad \hat{V}\left(\hat{X}\right) = \frac{N^2(1-f)}{n} \frac{\sum_{i}^{n} M_i^2 \left(\overline{X}_i - \overline{\overline{x}}\right)^2}{n-1}$$

Para el estimador de la proporción y el total de clase tenemos:

$$V(\hat{P}) = (1-f)\cdot\frac{N^2}{nM^2}\frac{\sum_i^N M_i^2(P_i - P)}{N-1}, \quad \hat{V}(\hat{P}) = (1-f)\cdot\frac{N^2}{nM^2}\frac{\sum_i^n M_i^2(P_i - \overline{P})}{n-1}$$

$$V(\hat{A}) = \frac{N^2(1-f)}{n}\frac{\sum_i^N M_i^2(P_i - P)^2}{N-1}, \quad \hat{V}(\hat{A}) = \frac{N^2(1-f)}{n}\frac{\sum_i^n M_i^2(P_i - \overline{P})^2}{n-1}$$

Muestreo con reposición

Por ser un estimador de la razón, su varianza aproximada es:

$$V(\overline{\overline{x}}) = \frac{N^2}{nM^2}\frac{\sum_i^N M_i^2(\overline{X}_i - \overline{X})}{N}, \quad \hat{V}(\overline{\overline{x}})\frac{N^2}{nM^2}\frac{\sum_i^n M_i^2(\overline{X}_i - \overline{\overline{x}})^2}{n-1}$$

Para el estimador del total tendremos:

$$V(\hat{X}) = \frac{N^2}{n}\frac{\sum_i^N M_i^2(\overline{X}_i - \overline{X})^2}{N}, \quad \hat{V}(\hat{X}) = \frac{N^2}{n}\frac{\sum_i^n M_i^2(\overline{X}_i - \overline{\overline{x}})^2}{n-1}$$

Para el estimador de la proporción y el total de clase tenemos:

$$V(\hat{P}) = \frac{N^2}{nM^2}\frac{\sum_i^N M_i^2(P_i - P)}{N} \Rightarrow \hat{V}(\hat{P}) = \frac{N^2}{nM^2}\frac{\sum_i^n M_i^2(P_i - \overline{P})}{n-1},$$

$$V(\hat{A}) = \frac{N^2}{n}\frac{\sum_i^N M_i^2(P_i - P)^2}{N}, \quad \hat{V}(\hat{A}) = \frac{N^2}{n}\frac{\sum_i^n M_i^2(P_i - \overline{P})^2}{n-1}$$

MUESTREO UNIETÁPICO DE CONGLOMERADOS DE DISTINTO TAMAÑO CON PROBABILIDADES DESIGUALES

En este caso se utilizan los estimadores generales de Horvitz Thompson y Hansen Hurweitz.

Muestreo sin reposición

Consideramos una población de N conglomerados de tamaños desiguales M_i con $M = \sum_{i=1}^{N} M_i$.

En este caso se utilizará el *estimador general de Horwitz y Thompson*, que proporciona el *estimador lineal insesgado para el total* definido por:

$$\hat{X}_{HT} = \sum_{i=1}^{n}\frac{X_i}{\pi_i} = \sum_{i=1}^{n}\frac{M_i\overline{X}_i}{\pi_i}, \quad V(\hat{X}_{HT}) = \sum_{i=1}^{N}\frac{X_i^2}{\pi_i}(1-\pi_i) + \sum_{i\neq j}^{N}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j)$$

$$\hat{V}(\hat{X}_{HT}) = \sum_{i=1}^{n}\frac{X_i^2}{\pi_i^2}(1-\pi_i) + \sum_{i\neq j}^{n}\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\right)$$

Muestreo con reposición

Consideramos una población de N conglomerados de tamaños desiguales M_i con $M = \sum_{i=1}^{N} M_i$. En este caso se utilizará el *estimador general de Hansen y Hurwitz*, que proporciona el *estimador lineal insesgado para el total* definido por:

$$\hat{X}_{HH} = \sum_{i=1}^{n} \frac{X_i}{nP_i} = \sum_{i=1}^{n} \frac{M_i \overline{X}_i}{nP_i}, \quad V(\hat{X}_{HH}) = \frac{1}{n} \sum_{i=1}^{N} \left(\frac{X_i}{P_i} - X \right)^2 P_i, \quad \hat{V}(\hat{X}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{X_i}{P_i} - \hat{X}_{HH} \right)^2$$

$$\hat{\overline{X}}_{HH} = \frac{\hat{X}_{HH}}{M} \Rightarrow V(\hat{\overline{X}}_{HH}) = V\left(\frac{\hat{X}_{HH}}{M}\right) = \frac{1}{M^2} V(\hat{X}_{HH}) \Rightarrow \hat{V}(\hat{\overline{X}}_{HH}) = \frac{1}{M^2} \hat{V}(\hat{X}_{HH})$$

Probabilidades proporcionales a los tamaños

Muestreo sin reposición

El *estimador lineal insesgado de Horwitz y Thompson para el total* será:

$$\hat{X}_{HT} = \sum_{i=1}^{n} \frac{X_i}{\pi_i} = \sum_{i=1}^{n} \frac{M_i \overline{X}_i}{\pi_i} = \sum_{i=1}^{n} \frac{M_i \overline{X}_i}{n \frac{M_i}{M}} = M \frac{1}{n} \sum_{i=1}^{n} \overline{X}_i = M \overline{\overline{x}}$$

El *estimador lineal insesgado de Horwitz y Thompson para la media* será:

$$\hat{\overline{X}} = \frac{\hat{X}_{HT}}{M} = \frac{M \overline{\overline{x}}}{M} = \overline{\overline{x}}$$

Se observa que las expresiones de los estimadores lineales insesgados para la media y el total en el caso de probabilidades desiguales proporcionales a los tamaños de los conglomerados coinciden con sus expresiones para probabilidades iguales.

Muestreo con reposición

Como siempre, los estimadores son los mismos que para el caso sin reposición. Las *varianzas y su estimación en el caso de probabilidades proporcionales a los tamaños con reposición* valdrán:

$$V(\hat{X}_{HH}) = \frac{M}{n} \sum_{i=1}^{N} M_i \left(\overline{X}_i - \overline{X} \right), \quad \hat{V}(\hat{X}_{HH}) = \frac{M^2}{n(n-1)} \sum_{i=1}^{n} \left(\overline{X}_i - \overline{\overline{x}} \right)^2$$

$$V(\hat{\overline{X}}_{HH}) = \frac{1}{nM} \sum_{i=1}^{N} M_i \left(\overline{X}_i - \overline{X} \right)^2, \quad \hat{V}(\hat{X}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\overline{X}_i - \overline{\overline{x}} \right)^2$$

Las *fórmulas para proporciones y totales de clase* se obtienen sustituyendo $\overline{X}_i = P_i, \quad \overline{X} = P, \quad \overline{\overline{x}} = \overline{P}$. Esto es válido tanto en general como en probabilidades proporcionales a los tamaños, y tanto con reposición como sin reposición.

TAMAÑO DE LA MUESTRA

La peculiaridad en muestreo por conglomerados monoetápico es la forma de la función de coste. Si consideramos la función de coste $C = c_o\sqrt{n} + c_1 n + c_2 \cdot n \cdot \overline{M}$, podemos determinar los pares $(n, \overline{M})$ que, para C prefijado, minimizan la varianza del estimador de la media $V(\bar{\bar{x}})$. También podemos determinar los pares $(n, \overline{M})$ que, para $V(\bar{\bar{x}})$ prefijada, minimizan la función de coste C.

El primer término $c_o\sqrt{n}$ de la función de coste representa el *coste de viaje entre los conglomerados*, y se toma así porque se ha demostrado empíricamente que el coste de viaje entre *n* conglomerados varía aproximadamente proporcional a su raíz cuadrada.

El segundo término $c_1 n$ de la función de coste representa el *coste de selección de los n conglomerados de la muestra*, siendo c_1 el coste unitario de selección de un conglomerado muestral.

El tercer término $c_2 \cdot n \cdot \overline{M}$ representa el *coste relativo a las* $n \cdot \overline{M}$ *unidades elementales de la muestra*, siendo c_2 el coste unitario de selección de una unidad elemental que suele estar formado principalmente por el coste de entrevista y el coste de desplazamiento entre las unidades elementales dentro del mismo conglomerado.

El término $C = c_o\sqrt{n} + c_2 \cdot n \cdot \overline{M}$ suele denominarse *coste de campo*.

La determinación de *n* y $\overline{M}$ óptimos lleva al planteamiento del problema de Lagrange con una restricción:

$$\begin{cases} Min V(\bar{\bar{x}}) = Min\left[(1-f)\dfrac{S^2}{n\overline{M}}(1-(\overline{M}-1)\delta)\right] \\ C = c_o\sqrt{n} + c_1 n + c_2 \cdot n \cdot \overline{M} \end{cases}$$

El problema alternativo es la determinación de *n* y $\overline{M}$ óptimos mediante el planteamiento del problema de Lagrange con una restricción:

$$\begin{cases} Min\, C = Min(c_o\sqrt{n} + c_1 n + c_2 \cdot n \cdot \overline{M}) \\ V(\bar{\bar{x}}) = (1-f)\dfrac{S^2}{n\overline{M}}(1-(\overline{M}-1)\delta) \end{cases}$$

También se utiliza para la varianza la expresión $V(\bar{\bar{x}}) = (1-f)\dfrac{S_b^2}{n\overline{M}}$.

PROBLEMAS RESUELTOS

7.1. Se trata de estudiar una población de 1000 cajas de tornillos todas ellas con 40 unidades cada una. Para ello se extrae una muestra sin reposición de 20 cajas, dentro de la cual nueve cajas no tienen tornillos defectuosos, ocho cajas tienen un tornillo defectuoso, y tres cajas tienen dos tornillos defectuosos. Se pide:

1) Estimar el número total de tornillos defectuosos en la población y sus errores absoluto y relativo de muestreo. Realizar la estimación por intervalos al 99% ($F^{-1}(0{,}995)=2{,}57$).
2) Resolver el problema con reposición y comparar los resultados con los del punto (a).

Tenemos como datos $N = 1000$, $\overline{M} = 40$ y $n = 20$. El total de piezas defectuosas puede estimarse como sigue:

$$\hat{A} = N\overline{M}\hat{P} = N\overline{M}\left(\frac{1}{n}\sum_{i=1}^{n} P_i\right) = 40\,000\underbrace{\frac{1}{20}\left(9\frac{0}{40}+8\frac{1}{40}+3\frac{2}{40}\right)}_{\hat{P}=0{,}0175} = 700$$

Para calcular la estimación de la varianza, se realiza el cuadro del análisis de la varianza muestral considerando 20 variables, desde L1 a L20, una para cada caja en la muestra. Cada variable tiene tantos unos como tornillos defectuosos hay en la caja. Se elige *Análisis de la varianza de un factor* en *Análisis de datos* del menú *Herramientas,* y se rellena su pantalla de entrada como se indica en la Figura 7-1. Los resultados se ven en la Figura 7-2. La varianza es:

$$\hat{V}(\hat{A}) = \left(N\overline{M}\right)^2\hat{V}(\hat{P}) = \left(N\overline{M}\right)^2(1-f)\frac{\hat{S}_b^2}{n\overline{M}} = 40000^2\left(1-\frac{20}{1000}\right)\frac{0{,}0134}{800} = 26305{,}26$$

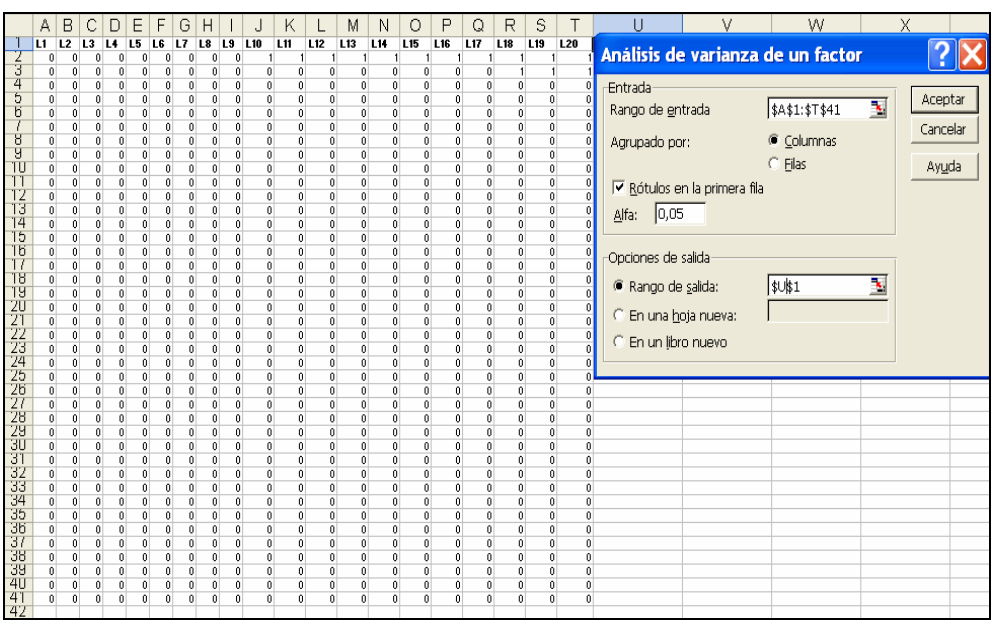


Figura 7-1

	U	V	W	X	Y	Z	AA
1	Análisis de varianza de un factor						
2							
3	RESUMEN						
4	*Grupos*	*Cuenta*	*Suma*	*Promedio*	*Varianza*		
5	L1	40	0	0	0		
6	L2	40	0	0	0		
7	L3	40	0	0	0		
8	L4	40	0	0	0		
9	L5	40	0	0	0		
10	L6	40	0	0	0		
11	L7	40	0	0	0		
12	L8	40	0	0	0		
13	L9	40	0	0	0		
14	L10	40	1	0,025	0,025		
15	L11	40	1	0,025	0,025		
16	L12	40	1	0,025	0,025		
17	L13	40	1	0,025	0,025		
18	L14	40	1	0,025	0,025		
19	L15	40	1	0,025	0,025		
20	L16	40	1	0,025	0,025		
21	L17	40	1	0,025	0,025		
22	L18	40	2	0,05	0,048717949		
23	L19	40	2	0,05	0,048717949		
24	L20	40	2	0,05	0,048717949		
25							
26							
27	ANÁLISIS DE VARIANZA						
28	*Origen de las variaciones*	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
29	Entre grupos	0,255	19	**0,013421053**	0,775438596	0,738076009	1,59987934
30	Dentro de los grupos	13,5	780	0,017307692			
31							
32	Total	13,755	799				
33							
34				VARIANZA ESTIMADOR TOTAL=	26305,26388		
35							

Figura 7-2

El error relativo de muestreo para el estimador del total será:

$$\hat{C}v(\hat{A}) = \frac{\sqrt{\hat{V}(\hat{A})}}{\hat{A}} = \frac{\sqrt{26305{,}26}}{700} = 0{,}2317 \quad (23{,}17\%)$$

La estimación por intervalos suponiendo normalidad en la población es:

$$\hat{A} \pm \lambda_\alpha \hat{\sigma}(\hat{A}) = 700 \pm 2{,}57\sqrt{26305{,}26} = \left[283{,}2,\ 1116{,}8\right]$$

La estimación por intervalos sin normalidad en la población es:

$$\hat{A} \pm \frac{\hat{\sigma}(\hat{A})}{\sqrt{\alpha}} = 700 \pm \sqrt{\frac{26305{,}26}{0{,}01}} = \left[-921{,}9,\ 2321{,}9\right]$$

Si consideramos muestreo con reposición, tenemos:

$$\hat{V}(\hat{A}) = \left(N\overline{M}\right)^2 \hat{V}(\hat{P}) = \left(N\overline{M}\right)^2 \frac{\hat{S}_b^2}{n\overline{M}} = \frac{26305{,}26}{1-f} = \frac{26305{,}26}{1-\frac{20}{1000}} = 26842{,}1$$

$$\hat{C}v(\hat{A}) = \frac{\sqrt{\hat{V}(\hat{A})}}{\hat{A}} = \frac{\sqrt{26842{,}1}}{700} = 0{,}234 \quad (23{,}4\%)$$

La estimación por intervalos suponiendo normalidad en la población es:

$$\hat{A} \pm \lambda_\alpha \hat{\sigma}(\hat{A}) = 700 \pm 2{,}57\sqrt{26842{,}1} = \left[279,\ 1121\right]$$

La estimación por intervalos sin normalidad en la población es:

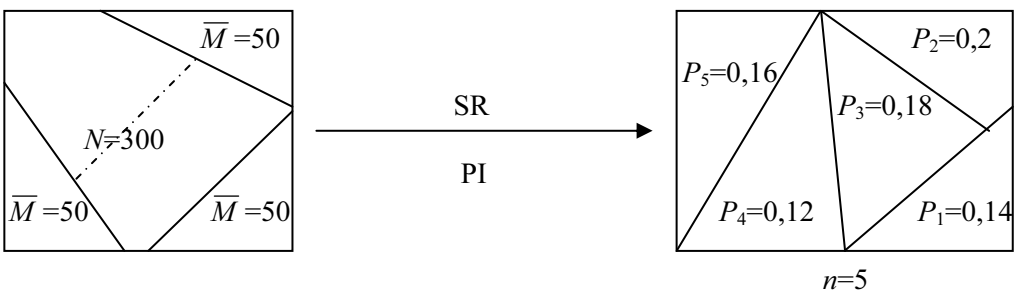
$$\hat{A} \pm \frac{\hat{\sigma}(\hat{A})}{\sqrt{\alpha}} = 700 \pm \sqrt{\frac{26842{,}1}{0{,}01}} = \left[-938{,}35,\ 2338{,}35\right]$$

Se observa que los errores de muestreo estimados son ligeramente superiores en muestreo con reposición. Además, como es natural, los intervalos de confianza son más anchos (o sea, peores) en muestreo con reposición. La ganancia en precisión es (26842,1 / 26305,26 – 1)100 = 2%, que es una cantidad pequeña.

7.2. En una región hay 300 granjas de 50 animales diversos cada una. Se obtiene una muestra de n=5 granjas sin reposición y probabilidades iguales. Las proporciones de animales enfermos en cada una de las granjas son 0,14, 0,20, 0,18, 0,12, 0,16. Se pide:

Estimar la proporción y el total de animales enfermos en la región y sus errores absoluto y relativo de muestreo. Realizar las mismas estimaciones para muestreo con reposición. Comentar los resultados.

Podemos realizar el esquema siguiente para el problema.



SR significa sin reposición y PI probabilidades iguales.

Estamos en un caso de muestreo monoetápico de conglomerados del mismo tamaño. Se tiene:

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} P_i = \frac{1}{5}(0{,}14 + 0{,}20 + 0{,}18 + 0{,}12 + 0{,}16) = 0{,}16$$

$$\hat{A} = N\overline{M}\hat{P} = 300 \cdot 50 \cdot 0{,}16 = 2400$$

$$\hat{V}(\hat{A}) = \left(N\overline{M}\right)^2 \hat{V}(\hat{P}) = \left(N\overline{M}\right)^2 (1-f)\frac{\hat{S}_b^2}{n\overline{M}} = \left(N\overline{M}\right)^2 (1-f)\frac{1}{n(n-1)}\sum_{i=1}^{n}(P_i - \hat{P})^2 =$$

$$\left(300 \cdot 50\right)^2 \left(1 - \frac{5}{300}\right)\frac{(0{,}14-0{,}16)^2 + (0{,}20-0{,}16)^2 + (0{,}18-0{,}16)^2 + (0{,}12-0{,}16)^2 + (0{,}16-0{,}16)^2}{5(5-1)} = 45000$$

$$\hat{V}(\hat{P}) = \frac{1}{N^2\overline{M}^2}\hat{V}(\hat{A}) = \frac{45000}{300^2 50^2} = 0{,}0002$$

$$\hat{C}v(\hat{P}) = \hat{C}v(\hat{A}) = \frac{\sqrt{\hat{V}(\hat{A})}}{\hat{A}} = \frac{\sqrt{45000}}{2400} = 0{,}088 \quad (8{,}8\%)$$

Se estima que en la región hay un 16% de animales enfermos y un total de 2400 animales enfermos, con un error de muestreo del 8,8%.

Ahora estimaremos los errores absoluto y relativo de muestreo del total de clase y de la proporción considerando muestreo con reposición. Tenemos:

$$\hat{V}(\hat{A}) = \left(N\overline{M}\right)^2 \hat{V}(\hat{P}) = \left(N\overline{M}\right)^2 \frac{\hat{S}_b^2}{n\overline{M}} = \left(N\overline{M}\right)^2 \frac{1}{n(n-1)} \sum_{i=1}^{n} (P_i - \hat{P})^2 = 45762{,}7$$

$$\hat{V}(\hat{P}) = \frac{1}{N^2 \overline{M}^2} \hat{V}(\hat{A}) = \frac{45762{,}7}{300^2 50^2} = 0{,}000203389$$

$$\hat{C}v(\hat{P}) = \hat{C}v(\hat{A}) = \frac{\sqrt{\hat{V}(\hat{A})}}{\hat{A}} = \frac{\sqrt{45762{,}7}}{2400} = 0{,}089 \quad (8{,}9\%)$$

Se observa que los errores de muestreo son ligeramente mayores en el caso de reposición.

7.3. En un proceso electoral se toma una muestra aleatoria de 10 urnas, el número de votantes y sus papeletas favorables a un determinado partido son:

Número de votantes	4	2	6	1	5	3	3	8	1	4
Papeletas favorables	2	1	4	1	2	1	2	5	0	3

Suponiendo muestreo con reposición, estimar la proporción de votos favorables a ese partido en toda la población y su error de muestreo.

Vamos a considerar las urnas como conglomerados, siendo las unidades elementales las papeletas introducidos en ellas. Por tanto, los números de papeletas en las distintas urnas serán los tamaños de los conglomerados M_i. Se considera la clase A de los votantes que votan a favor del partido en cuestión. Por tanto, las papeletas favorables al partido en cada urna serán los valores A_i.

Ya que los conglomerados son de distinto tamaño, para estimar la proporción del total de votantes de la población que votan al partido utilizaremos el estimador de la razón de A a M siguiente:

$$\hat{P} = \frac{\sum_{i=1}^{10} A_i}{\sum_{i=1}^{10} M_i} = \frac{21}{37} = 0{,}57$$

Para estimar la varianza de la proporción con reposición utilizamos el estimador de la varianza del estimador de la razón:

$$\hat{V}(\hat{P}) = \frac{1}{n\overline{M}^2}(\hat{S}_A^2 + \hat{R}^2 S_M^2 - 2\hat{R}\hat{S}_{AM}) = \frac{1}{n\overline{M}^2(n-1)}(\sum_{i=1}^{10} A_i^2 + \hat{R}^2 \sum_{i=1}^{10} M_i^2 - 2\hat{R}\sum_{i=1}^{10} A_i M_i)$$

$$= \frac{1}{10 \cdot 3{,}7^2 \cdot (10-1)}(65 + 0{,}57^2 \cdot 181 - 2 \cdot 0{,}57 \cdot 106) = 0{,}00242$$

El error de muestreo estimado será $\hat{\sigma}(\hat{P}) = \sqrt{\hat{V}(\hat{P})} = \sqrt{0{,}00242} = 0{,}049$.

7.4. Se trata de estudiar la superficie de una región montañosa dedicada a la plantación de pinos. La región, que tiene un total de 25000 km^2, se divide en 100 zonas disjuntas lo más similares entre sí de tal forma que cada zona contiene plantas de todas las clases que crecen en la región. Se extrae una muestra de 10 zonas con reemplazamiento y con probabilidades proporcionales a sus superficies. Las proporciones de superficie total dedicadas a la plantación de pinos en cada una de las zonas de la muestra son:

0,05, 0,25, 0,10, 0,30, 0,15, 0,25, 0,35, 0,25, 0,10 y 0,20

Se pide un estimador insesgado de la superficie total de la región dedicada a la plantación de pinos, su error relativo y un intervalo de confianza al nivel $\alpha = 0{,}05$.

Sea Mi = Superficie de la zona i-ésima
Sea Xi = Superficie dedicada a la plantación de pinos

$$\hat{X}_{HH} = \sum_{i=1}^{n} \frac{X_i}{nP_i} = \sum_{i=1}^{n} \frac{X_i}{n\frac{M_i}{M}} = \frac{M}{n}\sum_{i=1}^{n} \frac{X_i}{M_i} = \frac{2500}{10}(0{,}05 + 0{,}25 + \cdots + 0{,}20) = 5000$$

$$\hat{V}(\hat{X}_{HH}) = \frac{\sum_{i=1}^{n}\left(\frac{X_i}{P_i} - \hat{X}_{HH}\right)^2}{n(n-1)} = \frac{\sum_{i=1}^{n}\left(\frac{X_i}{M_i/M} - \hat{X}_{HH}\right)^2}{n(n-1)} = \frac{\sum_{i=1}^{n}\left(M\frac{X_i}{M_i} - \hat{X}_{HH}\right)^2}{n(n-1)} =$$

$$\frac{(25000 \cdot 0{,}05 - 5000)^2 + (25000 \cdot 0{,}25 - 5000)^2 + \cdots + (25000 \cdot 0{,}20 - 5000)^2}{10(10-1)} = 590278$$

$$\hat{C}v(\hat{X}) = \frac{\sqrt{\hat{V}(\hat{X})}}{\hat{X}} = \frac{\sqrt{590278}}{5000} = 0{,}15 \quad (15\%)$$

La estimación por intervalos suponiendo normalidad en la población es:

$$\hat{X} \pm \lambda_\alpha \hat{\sigma}(\hat{X}) = 5000 \pm 2\sqrt{590278} = [3464, 6536]$$

La estimación por intervalos sin normalidad en la población es:

$$\hat{X} \pm \frac{\hat{\sigma}(\hat{X})}{\sqrt{\alpha}} = 5000 \pm \sqrt{\frac{590278}{0{,}05}} = [1564, 8346]$$

7.5. Una gran empresa tiene sus inventarios de equipo listados separadamente en 15 departamentos. Se selecciona una muestra de tres departamentos con reposición y probabilidades proporcionales al número de artículos de equipo en cada departamento. La tabla siguiente presenta el número de artículos de equipo NA en cada departamento D.

D	*NA*	*D*	*NA*	*D*	*NA*	*D*	*NA*	*D*	*NA*
1	12	4	40	7	18	10	22	13	16
2	9	5	35	8	10	11	22	14	33
3	27	6	15	9	31	12	19	15	6

1) Suponiendo que los tres departamentos seleccionados (que serán los de mayor probabilidad) tienen cada uno 2 artículos impropiamente identificados, estimar el número total de artículos impropiamente identificados en la empresa y su error relativo de muestreo.

2) Estimar por intervalos al 95% la media de artículos propiamente identificados, sabiendo que los tres departamentos seleccionados tienen respectivamente 4, 5 y 6 artículos impropiamente identificados.

Como se selecciona la muestra de tres departamentos con probabilidades proporcionales al número de artículos de equipo en cada departamento, los tres departamentos seleccionados para la muestra serán el 4, el 5 y el 14, ya que son los que van a tener mayor probabilidad de selección (por tener el mayor número de artículos).

Al ser la selección con probabilidades proporcionales a los tamaños se tiene que:

$$P_i = \frac{M_i}{M} \Rightarrow P_1 = \frac{40}{315},\ P_2 = \frac{35}{315} \text{ y } P_3 = \frac{33}{315}$$

Como el muestreo es con reposición, el estimador insesgado del total de la clase de los artículos impropiamene clasificados vendrá dado por la fórmula de Hansen y Hurwitz.

$$\hat{\hat{A}}_{HH} = M\hat{\hat{P}}_{HH} = \frac{1}{n}\sum_i^n \frac{M_i\hat{P}_i}{P_i} = \frac{1}{n}\sum_i^n \frac{M_i\hat{P}_i}{M_i/M} = \frac{M}{n}\sum_i^n \hat{P}_i = \frac{315}{3}\left(\frac{2}{40}+\frac{2}{35}+\frac{2}{33}\right) \cong 18$$

$\hat{P}_i$ = proporción muestral en el conglomerado *i*-ésimo

Como estamos en muestreo monoetápico con reposición y probabilidades desiguales proporcionales a los tamaños, utilizamos para estimar la varianza el estimador:

$$\hat{V}\left(\hat{A}\right) = \frac{\sum_i^n \left(\frac{A_i}{P_i} - \hat{A}\right)^2}{n(n-1)} = \frac{\sum_i^n \left(\frac{M_i P_i}{P_i} - M\hat{P}\right)^2}{n(n-1)} = \frac{M^2\sum_i^n \left(P_i - \hat{P}\right)^2}{n(n-1)} =$$

$$\frac{315^2}{3\cdot 2}\left[\left(\frac{2}{40}-\frac{18}{315}\right)^2 + \left(\frac{2}{35}-\frac{18}{315}\right)^2 + \left(\frac{2}{33}-\frac{18}{315}\right)^2\right] = 1{,}04209$$

Para estimar la proporción de artículos propiamente identificados observamos que los tres departamentos seleccionados para la muestra (el 4, el 5 y el 14) tienen 36, 30 y 27 artículos propiamente identificados, respectivamente. El estimador será el siguiente:

$$\hat{P} = \frac{1}{n}\sum_{i}^{n}\frac{\frac{M_i}{M}P_i}{P_i} = \frac{1}{n}\sum_{i}^{n}\frac{\frac{M_i}{M}P_i}{M_i/M} = \frac{1}{n}\sum_{i}^{n}\hat{P}_i = \frac{1}{3}\left(\frac{36}{40}+\frac{30}{35}+\frac{27}{33}\right) = 0{,}858$$

$$\hat{V}\left(\hat{P}\right) = \frac{1}{M^2}\hat{V}\left(\hat{A}\right) = \frac{\sum_{i}^{n}\left(P_i-\hat{P}\right)^2}{n(n-1)} = \frac{1}{3\cdot 2}\left[\left(\frac{36}{40}-0{,}858\right)^2+\left(\frac{30}{35}-0{,}858\right)^2+\left(\frac{27}{33}-0{,}858\right)^2\right] = 0{,}000558$$

El intervalo de confianza al 95%, suponiendo normalidad, será:

$$\hat{P} \pm \lambda_\alpha\sqrt{\hat{V}(\hat{P})} = 0{,}858 \pm 1{,}96\sqrt{0{,}000558} = [0.8117,\ 0.9043]$$

7.6. Un fabricante de sierras quiere estimar el costo de reparación promedio mensual para las sierras que ha vendido a ciertas industrias. El fabricante no puede obtener un costo de reparación por sierra, pero puede obtener la cantidad total gastada en reparación y el número de sierras que tiene cada industria. El fabricante decide seleccionar una muestra aleatoria simple sin reposición de 20 industrias de entre las 96 a las que ofrece servicio. Los datos de gasto total mensual en reparaciones por industria y el número de sierras por industria se presentan en la tabla siguiente:

Indus.	*Nº de sierras*	*Costo total de reparaciones mensual*	*Indus.*	*Nº de sierras*	*Costo total de reparaciones mensual*
1	3	50	11	8	140
2	7	110	12	6	130
3	11	230	13	3	70
4	9	140	14	2	50
5	2	60	15	1	10
6	12	280	16	4	60
7	14	240	17	12	280
8	3	45	18	6	150
9	5	60	19	5	110
10	9	230	20	8	120

1) Estimar el costo promedio de reparación mensual por sierra y su error de muestreo.
2) Estimar la cantidad gastada por las 96 industrias en la reparación de sierras y su error de muestreo.
3) Después de verificar sus registros de ventas, el fabricante se percata de que ha vendido un total de 710 sierras a esas industrias. Usando esta información adicional, estimar la cantidad total gastada en reparación de sierras para estas industrias y su error de muestreo.
4) El mismo fabricante quiere estimar el coste de reparación promedio por sierra para el mes siguiente: ¿cuántos conglomerados debe seleccionar en la muestra si quiere que su error de muestreo sea inferior a una unidad?

Consideramos las industrias como conglomerados (N = 96). Se extrae una muestra de 20 conglomerados (n = 20) siendo las unidades elementales el número de sierras M_i de cada industria. El coste promedio de reparación de sierra se estimará como la razón entre el coste total de reparación por industria y el número de sierras por industria. Como los conglomerados son de tamaños desiguales tenemos:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i} = \frac{50+110+\cdots+120}{3+7+\cdots+8} = \frac{2565}{130} = 19{,}73$$

$$\hat{V}(\bar{\bar{x}}) = \frac{1-f}{n\bar{M}^2}(\hat{S}_x^2 + \hat{R}^2 S_M^2 - 2\hat{R}\hat{S}_{xm}) = \frac{1-f}{n\bar{M}^2(n-1)}(\sum_{i=1}^{10} X_i^2 + \hat{R}^2 \sum_{i=1}^{10} M_i^2 - 2\hat{R}\sum_{i=1}^{10} X_i M_i) =$$

$$\frac{1-\frac{20}{96}}{20\cdot\left(\frac{130}{20}\right)^2\cdot(20-1)}(460225+19{,}73^2\cdot 1188-2\cdot 19{,}73\cdot 22285) = 0{,}7905 \Rightarrow \hat{\sigma}(\bar{\bar{x}}) = 0{,}89$$

Para estimar el coste total en reparación de sierras en las industrias tomamos:

$$\hat{X} = \frac{N}{n}\sum_{i=1}^{n} X_i = \frac{96}{20}2565 = 12312$$

$$\hat{V}(\hat{X}) = N^2\,\frac{1-f}{n}\,\frac{\sum_{i=1}^{n}\left(X_i-\bar{\bar{x}}\right)^2}{n-1} = \frac{N^2(1-f)}{n(n-1)}(\sum_{i=1}^{n} X_i{}^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}) =$$

$$\frac{96^2(1-\frac{20}{96})}{20(20-1)}(460225-\frac{(2565)^2}{20}) = 25200516 \Rightarrow \hat{\sigma}(\hat{X}) = 1587{,}467$$

Ahora conocemos M = 710 y queremos estimar la cantidad total gastada para reparación de sierras en las industrias. Utilizaremos el estimador del total basado en la razón definido como:

$$\hat{X} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i}\cdot M = \frac{2565}{130}\cdot 710 = 14008{,}846$$

$$\hat{V}(\hat{X}) = N^2\frac{1-f}{n}(\hat{S}_x^2 + \hat{R}^2 S_M^2 - 2\hat{R}\hat{S}_{xm}) = \frac{N^2(1-f)}{n(n-1)}(\sum_{i=1}^{10} X_i^2 + \hat{R}^2\sum_{i=1}^{10} M_i^2 - 2\hat{R}\sum_{i=1}^{10} X_i M_i)$$

$$= \frac{96^2\left(1-\frac{20}{96}\right)}{20\cdot(20-1)}(460225+19{,}73^2\cdot 1188-2\cdot 19{,}73\cdot 22285) = 308467{,}24 \Rightarrow \hat{\sigma}(\hat{X}) = 555{,}4$$

El número n de conglomerados a seleccionar en la muestra si se quiere un error de muestreo inferior a una unidad al estimar el coste de reparación promedio por sierra para el mes siguiente se obtiene despejando n en la expresión:

$$\hat{V}(\bar{\bar{x}}) = \frac{1-\frac{n}{96}}{n\cdot\left(\frac{710}{96}\right)^2}\frac{16066002}{19} < 1 \Rightarrow n > 14$$

7.7. Un sociólogo quiere estimar el ingreso promedio por persona en una ciudad pequeña en la que no está disponible una lista de residentes. Par ello, se divide la ciudad en 415 bloques rectangulares de residentes sobre un mapa y se realizan entrevistas en 25 bloques. Se pregunta a los residentes de cada bloque por su ingreso total. Se obtienen los siguientes resultados:

Conglomerado	Número de residentes (Mi)	Ingreso total por conglomerado (Xi)
1	8	96000
2	12	121000
3	4	42000
4	5	65000
5	6	52000
6	6	40000
7	7	75000
8	5	65000
9	8	45000
10	3	50000
11	2	85000
12	6	43000
13	5	54000
14	10	49000
15	9	53000
16	3	50000
17	6	32000
18	5	22000
19	5	45000
20	4	37000
21	6	51000
22	8	30000
23	7	39000
24	3	47000
25	8	41000
SUMA→	151	1329000

1) Estimar el ingreso promedio por persona en la ciudad y establecer un límite para el error de estimación.
2) Estimar el ingreso total de todos los residentes de la ciudad y establecer un límite para el error de estimación sabiendo que hay 2500 residentes en la ciudad.
3) Estimar el ingreso total de todos los residentes de la ciudad y establecer un límite para el error de estimación si se desconoce el número de residentes en la ciudad.

Consideramos los bloques rectangulares de residentes como conglomerados (N = 415). Se extrae una muestra de 25 conglomerados (n = 25), siendo las unidades elementales el número de residentes M_i de cada bloque.

El ingreso promedio por persona en la ciudad se estimará como la razón entre el ingreso total de los bloques y el número de residentes en los bloques. Como los conglomerados son de tamaños desiguales tenemos:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i} = \frac{1329000}{151} = 8801$$

$$\hat{V}(\bar{\bar{x}}) = \frac{1-f}{n\bar{M}^2}(\hat{S}_x^2 + \hat{R}^2 S_M^2 - 2\hat{R}\hat{S}_{xm}) = \frac{1-f}{n\bar{M}^2(n-1)}(\sum_{i=1}^{10} X_i^2 + \hat{R}^2 \sum_{i=1}^{10} M_i^2 - 2\hat{R}\sum_{i=1}^{10} X_i M_i) = 653785$$

El límite para el error de estimación al 95% será:

$$\bar{\bar{x}} \pm 2\sqrt{\hat{V}(\bar{\bar{x}})} = 8801 \pm 2\sqrt{653785} = 8801 \pm 1617$$

Para estimar el ingreso total de todos los residentes de la ciudad hacemos lo siguiente:

$$\hat{X} = M\bar{\bar{x}} = 2500(8801) = 22002500$$

El error de estimación se estima mediante:

$$\hat{V}(\hat{X}) = M^2 \hat{V}(\bar{\bar{x}}) = 2500^2 (653785)$$

El límite para el error de estimación al 95% será:

$$\hat{X} \pm 2\sqrt{\hat{V}(\hat{X})} = 22002500 \pm 4042848$$

Si no se conocen los residentes en la ciudad *M*, para estimar el ingreso total de todos los residentes de la ciudad utilizamos el estimador:

$$\hat{X} = \frac{N}{n}\sum_{i=1}^{n} X_i = \frac{415}{25}1329000 = 22061400$$

$$\hat{V}(\hat{X}) = N^2 \frac{1-f}{n}\frac{\sum_{i=1}^{n}\left(X_i - \bar{\bar{x}}\right)^2}{n-1} = \frac{N^2(1-f)}{n(n-1)}(\sum_{i=1}^{n} X_i^{\,2} - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}) =$$

$$\frac{415^2(1-\frac{25}{415})}{25(25-1)}(82039000000 - \frac{(1329000)^2}{25}) \Rightarrow \hat{\sigma}(\hat{X}) = 1752960$$

El límite para el error de estimación al 95% será:

$$\hat{X} \pm 2\sqrt{\hat{V}(\hat{X})} = 22061400 \pm 3505920$$

7.8. Un auditor desea muestrear los registros de ausencias por enfermedad de una gran empresa, para estimar el número promedio de días de ausencia por enfermedad por empleado en el cuatrimestre pasado. La empresa tiene ocho divisiones, con diferentes números de empleados por división. Ya que el número de días de ausencia por enfermedad dentro de cada división debe estar altamente correlacionado con el número de empleados, el auditor decide muestrear n = 3 divisiones con probabilidad proporcional al número de empleados. Mostrar cómo seleccionar la muestra si los respectivos números de empleados son 1200, 450, 2100, 860, 2840, 1910, 390, 3200.

Supóngase que el número total de días de ausencia por enfermedad registrados en las tres divisiones muestreadas durante el cuatrimestre pasado son, respectivamente, $X_1 = 4320$, $X_2 = 4160$, $X_3 = 5790$. Estimar el número promedio de días de ausencia por enfermedad requeridos por persona, de toda la empresa, y establecer un límite para el error de estimación.

Comenzamos listando el número de empleados y el intervalo acumulado para cada división.

División	Número de empleados	Intervalo acumulado
1	1200	1-1200
2	450	1201-1650
3	2100	1651-3750
4	860	3751-4610
5	2840	4611-7450
6	1910	7451-9360
7	390	9361-9750
8	3200	9751-12950
	12950	

Como se van a muestrear $n = 3$ divisiones, debemos seleccionar tres números aleatorios entre 00001 y 12500. Los números obtenidos mediante una función generadora de números aleatorios automatizada resultan ser 02011, 07972 y 10281. El primero pertenece al intervalo acumulado de la división 3, el segundo al de la división 6 y el tercero al de la división 8. Por lo tanto, la muestra estará formada por las divisiones 3, 6 y 8.

$$\hat{\bar{X}}_{HH} = \frac{1}{M}\sum_{i=1}^{n}\frac{X_i}{nP_i} = \sum_{i=1}^{n}\frac{\frac{X_i}{M}}{n\frac{M_i}{M}} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{M_i} = \frac{1}{3}\left(\frac{4220}{3100}+\frac{4160}{1910}+\frac{5790}{3200}\right) = 2{,}02$$

$$\hat{V}(\hat{\bar{X}}_{HH}) = \frac{1}{M^2}\frac{\sum_{i=1}^{n}\left(\frac{X_i}{P_i}-\hat{X}_{HH}\right)^2}{n(n-1)} = \frac{\sum_{i=1}^{n}\left(\frac{X_i}{MP_i}-\frac{\hat{X}_{HH}}{M}\right)^2}{n(n-1)} = \frac{\sum_{i=1}^{n}\left(\frac{X_i}{M\frac{M_i}{M}}-\hat{\bar{X}}_{HH}\right)^2}{n(n-1)} =$$

$$\frac{\sum_{i=1}^{n}\left(\frac{X_i}{M_i}-\hat{\bar{X}}_{HH}\right)^2}{n(n-1)} = \frac{\left(\frac{4220}{3100}-2{,}02\right)^2+\left(\frac{4160}{1910}-2{,}02\right)^2+\left(\frac{5790}{3200}-2{,}02\right)^2}{3(3-1)} = 0{,}0119$$

El límite para el error de estimación será $2\sqrt{0{,}0119} = 0{,}22$.

EJERCICIOS PROPUESTOS

7.1. De una población formada por N conglomerados se selecciona una muestra de tamaño n con un procedimiento mediante el cual se elige la primera unidad para la muestra con probabilidades desiguales P_i, y los $n - 1$ conglomerados restantes de la muestra se eligen con probabilidades iguales, realizándose todas las extracciones sin reposición. Se pide una estimación insesgada del total poblacional X y sus errores absoluto y relativo de muestreo siendo $N = 50$, $n = 4$, X_i el total del conglomerado i-ésimo y conociendo los siguientes datos de los conglomerados de la muestra:

P_i	0,026	0,017	0,022	0,013
X_i	100	80	120	60

7.2. En una población compuesta por 10 conglomerados de 100 elementos se toma una muestra monoetápica de n conglomerados. Por experiencias anteriores se sabe que el modelo de Smith $S^2_b = S^2\ \overline{M}_t$ se ajusta bien en la proximidad de $\overline{M} = 100$ y se conoce el valor de $S^2_b = 1173$. Se pide:

Calcular el valor de t y S^2_w en el supuesto de que $S^2_b / S^2 = 13{,}8$.

Formar la tabla poblacional del análisis de la varianza y hallar el coeficiente de correlación intraconglomerados.

7.3. Una industria está considerando la revisión de su política de jubilación y quiere estimar la proporción de empleados que apoyan la nueva política. La industria consiste de 87 plantas separadas localizadas en todo Estados Unidos. Ya que los resultados deben ser obtenidos rápidamente y con poco dinero, la industria decide usar muestreo por conglomerados, con cada planta como un conglomerado. Se selecciona una muestra irrestricta aleatoria de 15 plantas y se obtienen las opiniones de los empleados en estas plantas a través de un cuestionario. Los resultados se presentan en la tabla anexa. Estimar la proporción de empleados en la industria que apoyan la nueva política de jubilación y establecer un límite para el error de estimación.

Planta	Número de empleados	Número de empleados que apoyan la nueva política
1	51	42
2	62	53
3	49	40
4	73	45
5	101	63
6	48	31
7	65	38
8	49	30
9	73	57
10	61	45
11	58	51
12	52	29
13	65	46
14	49	37
15	55	42

7.4. El gerente de circulación de un periódico desea estimar el número promedio de ejemplares comprados por familia en determinada comunidad. Los costos de transporte de un hogar a otro son sustanciales. Es por eso por lo que se listan los 4000 hogares de la comunidad en 400 conglomerados geográficos de 10 hogares cada uno, y se selecciona una muestra irrestricta aleatoria de 4 conglomerados. Se realizan las entrevistas con los resultados que se muestran en la tabla anexa. Estimar el número promedio de periódicos por hogar en la comunidad y establecer un límite para el error de estimación.

Conglomerado	Número de periódicos										Total
1	1	2	1	3	3	2	1	4	1	1	19
2	1	3	2	2	3	1	4	1	1	2	20
3	2	1	1	1	1	3	2	1	3	1	16
4	1	1	3	2	1	5	1	2	3	1	20

7.5. Se diseña una encuesta económica para estimar la cantidad promedio gastada en servicios para el hogar en una ciudad. Ya que no se encuentra disponible una lista de hogares, se usa muestreo por conglomerados, con divisiones (barrios) formando los conglomerados. Se selecciona una muestra aleatoria de 20 barrios de la ciudad de un total de 60. Los entrevistadores obtienen el costo de los servicios de cada hogar dentro de los barrios seleccionados; los costos totales se muestran en la tala anexa. Estimar la cantidad promedio de gastos en servicios por hogar en la ciudad y establecer un límite para el error de estimación.

Barrio muestreado	Número de hogares	Cantidad total gastada en servicios
1	55	2210
2	60	2390
3	63	2430
4	58	2380
5	71	2760
6	78	3110
7	69	2780
8	58	2370
9	52	1990
10	71	2810
11	73	2930
12	64	2470
13	69	2830
14	58	2370
15	63	2390
16	75	2870
17	78	3210
18	51	2430
19	67	2730
20	70	2880

CAPÍTULO

8

MUESTREO BIETÁPICO DE CONGLOMERADOS

OBJETIVOS

1. Presentar el concepto de muestreo de conglomerados en dos etapas.

2. Analizar los estimadores y sus errores en muestreo bietápico de conglomerados del mismo tamaño con probabilidades iguales.

3. Analizar los estimadores y sus errores en muestreo bietápico de conglomerados del mismo tamaño con probabilidades iguales considerando todas las opciones posibles de reposición o no en ambas etapas.

4. Analizar los estimadores y sus errores en muestreo bietápico de conglomerados de distinto tamaño con probabilidades iguales.

5. Analizar los estimadores y sus errores en muestreo bietápico de conglomerados de distinto tamaño con probabilidades iguales considerando todas las opciones posibles de reposición o no en ambas etapas.

6. Estudiar el tamaño de la muestra en muestreo bietápico.

7. Analizar los estimadores y sus errores en muestreo bietápico de conglomerados con probabilidades desiguales y con reposición en primera etapa.

8. Analizar los estimadores y sus errores en muestreo bietápico de conglomerados con probabilidades desiguales y sin reposición en primera etapa.

9. Presentar el concepto de muestreo polietápico.

10. Analizar los estimadores y sus errores en muestreo polietápico.

11. Estudiar diseños polietápicos complejos.

12. Estudiar el muestreo bietápico con estratificación en primera etapa.

ÍNDICE

1. Muestreo bietápico de conglomerados. Estimadores para probabilidades iguales y conglomerados del mismo tamaño.

2. Varianzas y su estimación en muestreo bietápico con probabilidades iguales y conglomerados del mismo tamaño.

3. Muestreo bietápico de conglomerados de distinto tamaño y probabilidades iguales.

4. Tamaño de la muestra en muestreo bietápico.

5. Muestreo bietápico con probabilidades desiguales y con reposición en 1ª etapa. Estimadores, varianzas y su estimación.

6. Muestreo bietápico con probabilidades desiguales y sin reposición en 1ª etapa. Estimadores, varianzas y su estimación.

7. Muestreo polietápico.

8. Diseños complejos: Muestreo bietápico con estratificación en primera etapa.

9. Problemas resueltos.

10. Ejercicios propuestos.

MUESTREO BIETÁPICO DE CONGLOMERADOS. ESTIMADORES PARA PROBABILIDADES IGUALES Y CONGLOMERADOS DEL MISMO TAMAÑO

El muestreo bietápico de conglomerados es un tipo de muestreo en el que en una primera etapa se selecciona una muestra de n conglomerados de tamaños M_i, $i = 1, 2, ..., n$ y en una segunda etapa se selecciona, independientemente en cada conglomerado de la primera etapa, una submuestra de m_i unidades elementales de entre las Mi del conglomerado. En ambas etapas la selección puede ser con o sin reposición, pero en la segunda etapa suele usarse muestreo sin reposición. En la segunda etapa se puede utilizar cualquier tipo de muestreo de los ya estudiados, pero generalmente sin reposición y probabilidades iguales.

En el muestreo bietápico no es necesario utilizar todas las unidades elementales de los conglomerados seleccionados en primera etapa. Tampoco es necesario un marco de unidades elementales completo; basta con un marco más basto para conglomerados, y dentro de cada conglomerado basta con un submarco para el submuestreo en segunda etapa. De esta forma, a medida que se consideran etapas de submuestreo se utilizan submarcos más bastos, y por lo tanto más fáciles de conseguir y manejar, que los marcos completos de unidades elementales. Cuando hay un cierto grado de homogeneidad dentro de los conglomerados muestrales es absurdo seleccionar todas sus unidades elementales para la muestra. Bastará con elegir sólo algunas de ellas originándose el submuestreo. En el muestreo bietápico se necesitan menos recursos y el coste es menor, ya que sólo se visitan algunas de las unidades elementales de los conglomerados elegidos en primera etapa para la muestra. No obstante, en el muestreo bietápico la precisión es menor; los submarcos dentro de cada conglomerado pueden originar complicaciones al aumentar el número de etapas de submuestreo y aparecen fuentes de variación que complican los cálculos algebraicos (tantas fuentes como etapas tenga el muestreo). La primera fuente es debida a la selección de las unidades primarias y la fuente 2 es debida al submuestreo dentro de cada unidad primaria.

El muestreo bietápico también se denomina muestreo en dos etapas o muestreo con submuestreo (el submuestreo es la segunda etapa).

Un estimador insesgado de la media será, lógicamente, la media muestral de las medias muestrales derivadas del submuestreo dentro de cada conglomerado:

$$\bar{\bar{x}} = \frac{1}{n\bar{m}} \sum_{i}^{n} \sum_{j}^{\bar{m}} X_{ij} = \frac{1}{n} \sum_{i}^{n} \bar{x}_i$$

Para el total poblacional, proporción y total de clase, los estimadores insesgados son los siguientes:

$$\hat{X} = N\,\bar{M}\,\bar{\bar{x}} = \frac{N\bar{M}}{n} \sum_{i}^{n} \bar{x}_i, \quad \hat{P} = \frac{1}{n} \sum_{i}^{n} \hat{P}_i, \hat{A} = N\bar{M}\hat{P} = \frac{N\bar{M}}{n} \sum_{i}^{n} \hat{P}_i$$

VARIANZAS Y SU ESTIMACIÓN EN MUESTREO BIETÁPICO CON PROBABILIDADES IGUALES Y CONGLOMERADOS DEL MISMO TAMAÑO

Las expresiones para la varianzas de los estimadores en el muestreo bietápico dependerán de las fracciones de muestreo en ambas etapas y de la reposición. Tenemos:

Muestreo sin reposición en las dos etapas

$$V\left(\bar{\bar{x}}\right)=\left(1-f_1\right)\cdot\frac{S_b^2}{n\overline{M}}+\left(1-f_2\right)\cdot\frac{S_w^2}{n\overline{m}}$$

$$f_1=\frac{n}{N},\ f_2=\frac{\overline{m}}{\overline{M}},\ S_b^2=\overline{M}\frac{\sum_i^N\left(\bar{X}_i-\bar{\bar{X}}\right)}{N-1},\ S_w^2=\frac{\sum_j^{\overline{M}}\left(X_{ij}-\bar{X}_i\right)^2}{\left(\overline{M}-1\right)\cdot N}$$

$$V(\hat{X})=N^2V\left(\bar{\bar{x}}\right)=\left(1-f_1\right)\cdot\frac{N^2\overline{M}S_b^2}{n}+\left(1-f_2\right)\cdot\frac{N^2\overline{M}^2S_w^2}{n\overline{m}}$$

$$V\left(\hat{P}\right)=\left(1-f_1\right)\frac{\frac{1}{N-1}\sum_i^N\overline{M}\left(P_i-P\right)^2}{n\overline{M}}+\left(1-f_2\right)\frac{\frac{1}{N(\overline{M}-1)}\sum_{i=1}^N\overline{M}P_i\left(1-P_i\right)}{n\overline{m}}=$$

$$\left(1-f_1\right)\frac{\sum_i^N\left(P_i-P\right)^2}{n(N-1)}+\left(1-f_2\right)\frac{\sum_{i=1}^N\overline{M}P_i\left(1-P_i\right)}{n\overline{m}N(\overline{M}-1)}$$

$$V(\hat{A})=N^2\overline{M}^2V(\hat{P})=\left(1-f_1\right)\frac{N^2\overline{M}^2\sum_i^N\left(P_i-P\right)^2}{n(N-1)}+\left(1-f_2\right)\frac{N\overline{M}^3\sum_{i=1}^NP_i\left(1-P_i\right)}{n\overline{m}(\overline{M}-1)}$$

A partir de la tabla de descomposición del análisis de la varianza muestral, pueden realizarse las estimaciones de las varianzas. La citada tabla es la siguiente:

Fuente	*Grados libertad*	*Sumas de cuadrados*	*Cuadrados medios*	*Valores esperados*
"*Entre*"	$n-1$	$\overline{m}\sum_i^n\left(\bar{x}_i-\bar{\bar{x}}\right)^2$	$\hat{S}_b^2$	$\frac{\overline{m}}{\overline{M}}S_b^2+\left(1-f_2\right)S_w^2$
"*Dentro*"	$n\left(\overline{m}-1\right)$	$\sum_i^n\sum_j^{\overline{m}}\left(X_{ij}-\bar{x}_i\right)^2$	$\hat{S}_w^2$	S_w^2
Total	$n\overline{m}-1$	$\sum_i^n\sum_j^{\overline{m}}\left(X_{ij}-\bar{\bar{x}}\right)^2$	$\hat{S}^2$	S^2

Las estimaciones de las varianzas para las dos etapas sin reposición son las siguientes:

$$\hat{V}\left(\bar{\bar{x}}\right)=\left(1-f_1\right)\frac{\hat{S}_b^2}{n\overline{m}}+f_1\left(1-f_2\right)\frac{\hat{S}_w^2}{n\overline{m}},\ \text{y}\quad\hat{V}(\hat{X})=N^2\overline{M}^2\hat{V}(\bar{\bar{x}})$$

$$\hat{V}\left(\hat{P}\right)=\left(1-f_1\right)\cdot\frac{\sum_i^n\left(P_i-\overline{P}\right)^2}{n\left(n-1\right)}+f_1\left(1-f_2\right)\cdot\frac{\sum_i^nP_iQ_i}{n^2\left(\overline{m}-1\right)}$$

$$\hat{V}(\hat{X}) = N^2\,\overline{M}^2\hat{V}(\overline{\overline{x}}) \quad y \quad \hat{V}(\hat{A}) = N^2\,\overline{M}^2\hat{V}(\hat{P})$$

Si f_1 es muy pequeña, se toma $\hat{V}\left(\overline{\overline{x}}\right) = \left(1 - f_1\right)\cdot\dfrac{\hat{S}_b^2}{n\overline{m}}$.

Muestreo con reposición en las dos etapas

$$V\left(\overline{\overline{x}}\right) = \frac{\sigma_b^2}{n\overline{M}} + \frac{\sigma_w^2}{n\overline{m}}$$

$$V(\hat{X}) = V\left(N\overline{M}\overline{\overline{x}}\right) = \frac{N^2\,\overline{M}\sigma_b^2}{n} + \frac{N^2\,\overline{M}^2\sigma_w^2}{n\overline{m}}$$

$$V\left(\hat{P}\right) = \frac{\dfrac{1}{N}\sum_i^N \overline{M}\left(P_i - P\right)^2}{n\overline{M}} + \frac{\dfrac{1}{N\overline{M}}\sum_{i=1}^N \overline{M}P_i\left(1 - P_i\right)}{n\overline{m}} = \frac{\sum_i^N\left(P_i - P\right)^2}{nN} + \frac{\sum_{i=1}^N P_i\left(1 - P_i\right)}{n\overline{m}N}$$

$$V(\hat{A}) = N^2\,\overline{M}^2V(\hat{P}) = \frac{N\overline{M}^2\sum_i^N\left(P_i - P\right)^2}{n} + \frac{N\overline{M}^2\sum_{i=1}^N P_i\left(1 - P_i\right)}{n\overline{m}}$$

Las estimaciones de varianzas son:

$$\hat{V}\left(\overline{\overline{x}}\right) = \frac{\hat{S}_b^2}{n\overline{m}}, \quad \text{y} \quad \hat{V}(\hat{X}) = N^2\,\overline{M}^2\hat{V}(\overline{\overline{x}})$$

$$\hat{V}\left(\hat{P}\right) = \frac{\dfrac{\overline{m}}{n-1}\sum_i^n\left(P_i - \overline{P}\right)^2}{n\overline{m}} = \frac{\sum_i^n\left(P_i - \overline{P}\right)^2}{n(n-1)} \quad \text{y} \quad \hat{V}\left(\hat{A}\right) = N^2\,\overline{M}^2\,\frac{\sum_i^n\left(P_i - \overline{P}\right)^2}{n(n-1)}$$

Primera etapa con reposición y segunda sin reposición

$$V\left(\overline{\overline{x}}\right) = \frac{\sigma_b^2}{n\overline{M}} + (1 - f_2)\frac{S_w^2}{n\overline{m}}$$

$$V(\hat{X}) = V\left(N\overline{M}\overline{\overline{x}}\right) = \frac{N^2\,\overline{M}\sigma_b^2}{n} + (1 - f_2)\frac{N^2\,\overline{M}^2S_w^2}{n\overline{m}}$$

$$V\left(\hat{P}\right) = \frac{\dfrac{1}{N}\sum_i^N\overline{M}\left(P_i - P\right)^2}{n\overline{M}} + (1 - f_2)\frac{\dfrac{1}{N(\overline{M}-1)}\sum_{i=1}^N\overline{M}P_i\left(1 - P_i\right)}{n\overline{m}} = \frac{\sum_i^N\left(P_i - P\right)^2}{nN} + (1 - f_2)\frac{\sum_{i=1}^N\overline{M}P_i\left(1 - P_i\right)}{n\overline{m}N(\overline{M}-1)}$$

$$V(\hat{A}) = N^2\,\overline{M}^2V(\hat{P}) = \frac{N\overline{M}^2\sum_i^N\left(P_i - P\right)^2}{n} + (1 - f_2)\frac{N\overline{M}^3\sum_{i=1}^N P_i\left(1 - P_i\right)}{n\overline{m}(\overline{M}-1)}$$

Las estimaciones de varianzas son iguales que para reposición en las dos etapas:

$$\hat{V}\left(\bar{\bar{x}}\right)=\frac{\hat{S}_b^2}{n\overline{m}},\quad \text{y}\quad \hat{V}(\hat{X})=N^2\overline{M}^2\hat{V}(\bar{\bar{x}})$$

$$\hat{V}\left(\hat{P}\right)=\frac{\frac{\overline{m}}{n-1}\sum_i^n\left(P_i-\overline{P}\right)^2}{n\overline{m}}=\frac{\sum_i^n\left(P_i-\overline{P}\right)^2}{n(n-1)}\quad \text{y}\quad \hat{V}\left(\hat{A}\right)=N^2\overline{M}^2\frac{\sum_i^n\left(P_i-\overline{P}\right)^2}{n(n-1)}$$

Primera etapa sin reposición y segunda con reposición

$$V\left(\bar{\bar{x}}\right)=(1-f_1)\frac{S_b^2}{n\overline{M}}+\frac{\sigma_w^2}{n\overline{m}}$$

$$V(\hat{X})=V\left(N\overline{M}\bar{\bar{x}}\right)=(1-f_1)\frac{N^2\overline{M}S_b^2}{n}+\frac{N^2\overline{M}^2\sigma_w^2}{n\overline{m}}$$

$$V\left(\hat{P}\right)=(1-f_1)\frac{\frac{1}{N-1}\sum_i^N\overline{M}\left(P_i-P\right)^2}{n\overline{M}}+\frac{\frac{1}{N\overline{M}}\sum_{i=1}^N\overline{M}P_i\left(1-P_i\right)}{n\overline{m}}=(1-f_1)\frac{\sum_i^N\left(P_i-P\right)^2}{n(N-1)}+\frac{\sum_{i=1}^N P_i\left(1-P_i\right)}{n\overline{m}N}$$

$$V(\hat{A})=N^2\overline{M}^2V(\hat{P})=(1-f_1)\frac{N^2\overline{M}^2\sum_i^N\left(P_i-P\right)^2}{n(N-1)}+\frac{N\overline{M}^2\sum_{i=1}^N P_i\left(1-P_i\right)}{n\overline{m}}$$

Cuando la primera etapa es sin reposición y la segunda con reposición, las estimaciones de varianzas son:

$$\hat{V}\left(\bar{\bar{x}}\right)=\left(1-f_1\right)\frac{\hat{S}_b^2}{n\overline{m}}+f_1\frac{\hat{S}_w^2}{n\overline{m}}\ ,\quad \text{y}\quad \hat{V}(\hat{X})=N^2\overline{M}^2\hat{V}(\bar{\bar{x}})$$

$$V\left(\hat{P}\right)=(1-f_1)\frac{\sum_i^N\left(P_i-P\right)^2}{n(N-1)}+\frac{\sum_{i=1}^N P_i\left(1-P_i\right)}{n\overline{m}N}\quad \text{y}\quad V(\hat{A})=N^2\overline{M}^2V(\hat{P})$$

Para proporciones y totales de clase: $\hat{S}_b^2=\frac{\overline{m}}{n-1}\sum_i^n\left(P_i-\overline{P}\right)^2$ y $\hat{S}_w^2=\frac{\sum_{i=1}^n\overline{m}P_i\left(1-P_i\right)}{n\left(\overline{m}-1\right)}$.

MUESTREO BIETÁPICO DE CONGLOMERADOS DE DISTINTO TAMAÑO Y PROBABILIDADES IGUALES

Para probabilidades iguales se tiene: $\hat{X}=N\frac{1}{n}\sum_i^n M_i\bar{x}_i=\frac{N}{n}\sum_i^n M_i\bar{x}_i$.

Las dos etapas sin reposición

Las varianzas y sus estimaciones para las dos etapas sin reposición son las siguientes:

$$V\left(\hat{X}\right)=N^2\cdot\left(1-f_1\right)\frac{\sum_i^N\left(X_i-\overline{X}\right)^2}{n\left(N-1\right)}+\frac{N}{n}\sum_i^N M_i^2\cdot\left(1-f_{2i}\right)\cdot\frac{\sum_j^{M_i}\left(X_{ij}-\overline{X}_i\right)^2}{\left(M_i-1\right)m_i}$$

$$\hat{V}\left(\hat{X}\right)=\frac{N^2\left(1-f_1\right)}{n}\cdot\frac{\sum_i^n\left(\hat{X}_i-\hat{\overline{X}}_i\right)^2}{n-1}+\frac{N}{n}\sum_i^n\frac{M_i^2\left(1-f_{2i}\right)}{m_i}\cdot\frac{\sum_j^{m_i}\left(X_{ij}-\overline{x}_i\right)^2}{m_i-1}$$

$$\left(\hat{\overline{X}}_i=\frac{1}{n}\sum_i^n\hat{X}_i,\ \hat{X}_i=M_i\overline{x}\right)_i$$

Primera etapa sin reposición y segunda etapa con reposición

En este caso, las varianzas y sus estimaciones son las siguientes:

$$V\left(\hat{X}\right)=N^2\cdot\left(1-f_1\right)\frac{\sum_i^N\left(X_i-\overline{X}\right)^2}{n\left(N-1\right)}+\frac{N}{n}\sum_i^N\frac{M_i}{m_i}\sum_j^{M_i}\left(X_{ij}-\overline{X}_i\right)^2$$

$$\hat{V}\left(\hat{X}\right)=\frac{N^2\left(1-f_1\right)}{n}\cdot\frac{\sum_i^n\left(\hat{X}_i-\hat{\overline{X}}_i\right)^2}{n-1}+\frac{N}{n}\sum_i^n\frac{M_i^2}{m_i}\cdot\frac{\sum_j^{m_i}\left(X_{ij}-\overline{x}_i\right)^2}{m_i-1}$$

$$\left(\hat{\overline{X}}_i=\frac{1}{n}\sum_i^n\hat{X}_i,\quad\hat{X}_i=M_i\overline{x}_i\right)$$

Las dos etapas con reposición

En este caso, las varianzas y sus estimaciones son las siguientes:

$$V\left(\hat{X}\right)=\frac{N}{n}\cdot\sum_i^{N_i}\left(X_i-\overline{X}\right)^2+\frac{N}{n}\sum_i^N\frac{M_i}{m_i}\sum_j^{M_i}\left(X_{ij}-\overline{X}_i\right)^2$$

$$\hat{V}\left(\hat{X}\right)=\frac{N^2}{n}\cdot\frac{\sum_i^n\left(\hat{X}_i-\hat{\overline{X}}_i\right)^2}{n-1}\qquad\left(\hat{\overline{X}}_i=\frac{1}{n}\sum_i^n\hat{X}_i\ y\ \hat{X}_i=M_i\overline{x}_i\right)$$

Primera etapa con reposición y segunda sin reposición

En este caso, las varianzas y sus estimaciones son las siguientes:

$$V\left(\hat{X}\right)=\frac{N}{n}\sum_i^N\left(X_i-\overline{X}\right)^2+\frac{N}{n}\sum_i^N M_i^2\cdot\left(1-f_{2i}\right)\cdot\frac{\sum_j^{M_i}\left(X_{ij}-\overline{X}_i\right)^2}{\left(M_i-1\right)m_i}$$

$$\hat{V}\left(\hat{X}\right)=\frac{N^2}{n}\cdot\frac{\sum_i^n\left(\hat{X}_i-\hat{\overline{X}}_i\right)^2}{n-1}\qquad\left(\hat{\overline{X}}_i=\frac{1}{n}\sum_i^n\hat{X}_i\ y\ \hat{X}_i=M_i\overline{x}_i\right)$$

Para ***proporciones y totales de clase***: $\hat{\bar{X}}_i = \frac{1}{n}\sum_i^n M_i \hat{P}_i$ *y* $\hat{X}_i = M_i \hat{P}_i$

Los *estimadores para medias, proporciones y totales de clase en el muestreo bietápico con probabilidades iguales y conglomerados de distinto tamaño* son inmediatos:

$$\hat{\bar{X}} = \frac{\hat{X}}{M} = \frac{N}{n}\sum_i^n \frac{M_i}{M}\bar{x}_i\,,\ V(\hat{\bar{X}}) = \frac{1}{M^2}V(\hat{X}),\ \hat{V}(\hat{\bar{X}}) = \frac{1}{M^2}\hat{V}(\hat{X})$$

$$\hat{P} = \frac{N}{n}\sum_i^n \frac{M_i}{M}\hat{P}_i\,,\ \hat{A} = M\hat{P} = \frac{N}{n}\sum_i^n M_i\hat{P}_i$$

$\hat{P}_i$ = proporción muestral en el conglomerado i-ésimo

Las fórmulas para la varianza del total de clase y su estimación en el caso de *muestreo sin reposición en ambas etapas* son las siguientes:

$$V\left(\hat{A}\right) = (1-f_1)\frac{N^3 PQ}{n(N-1)} + \frac{N}{n}\sum_i^N M_i^3 \cdot (1-f_{2i}) \cdot \frac{P_i Q_i}{(M_i - 1)m_i}$$

$$\hat{V}\left(\hat{A}\right) = \frac{N^2(1-f_1)}{n} \cdot \frac{\sum_i^n \left(M_i\hat{P}_i - \frac{1}{n}\sum_{i=1}^n M_i\hat{P}_i\right)^2}{n-1} + \frac{N}{n}\sum_i^n M_i^2(1-f_{2i}) \cdot \frac{\hat{P}_i\hat{Q}_i}{m_i - 1}$$

Las fórmulas para la varianza del total de clase y su estimación en el caso de *muestreo sin reposición en primera etapa y con reposición en* segunda son las siguientes:

$$V\left(\hat{A}\right) = (1-f_1)\frac{N^3 PQ}{n(N-1)} + \frac{N}{n}\sum_i^N \frac{M_i^2}{m_i} P_i Q_i$$

$$\hat{V}\left(\hat{A}\right) = \frac{N^2(1-f_1)}{n} \cdot \frac{\sum_i^n \left(M_i\hat{P}_i - \frac{1}{n}\sum_{i=1}^n M_i\hat{P}_i\right)^2}{n-1} + \frac{N}{n}\sum_i^n M_i^2 \cdot \frac{\hat{P}_i\hat{Q}_i}{m_i - 1}$$

Las fórmulas para la varianza del total de clase y su estimación en el caso de *muestreo con reposición en ambas etapas* son las siguientes:

$$V\left(\hat{A}\right) = \frac{N^2}{n} PQ + \frac{N}{n}\sum_i^N \frac{M_i^2}{m_i} P_i Q_i$$

$$\hat{V}\left(\hat{A}\right) = \frac{N^2}{n} \cdot \frac{\sum_i^n \left(M_i\hat{P}_i - \frac{1}{n}\sum_{i=1}^n M_i\hat{P}_i\right)^2}{n-1}$$

Las fórmulas para la varianza del total de clase y su estimación en el caso de *muestreo con reposición en primera etapa y sin reposición en segunda* son las siguientes:

$$V\left(\hat{A}\right)=\frac{N^2}{n}PQ+\frac{N}{n}\sum_{i}^{N}M_i^3\cdot\left(1-f_{2i}\right)\cdot\frac{P_iQ_i}{\left(M_i-1\right)m_i}$$

$$\hat{V}\left(\hat{A}\right)=\frac{N^2}{n}\cdot\frac{\sum_{i}^{n}\left(M_i\hat{P}_i-\frac{1}{n}\sum_{i=1}^{n}M_i\hat{P}_i\right)^2}{n-1}$$

Para proporciones aplicamos $V(\hat{P})=\frac{1}{M^2}V(\hat{A})$ y $\hat{V}(\hat{P})=\frac{1}{M^2}\hat{V}(\hat{A})$.

TAMAÑO DE LA MUESTRA EN MUESTREO BIETÁPICO

Suele expresarse el coste total *C* mediante la *función general de costes* $f\left(n,\overline{M},\overline{m}\right)$ definida como:

$$C=c_o+c_1n^{a_1}+c_2\left(n\overline{M}\right)^{a_2}+c_3\left(n\overline{M}\right)^{a_3}$$

en donde c_0 representa un coste fijo que suele incluir, dependiendo de las encuestas, gastos de preparación técnica, gastos administrativos previos, cartografía, etc. Puede empezarse por suponer deducido el coste c_o del total *C*, para no preocuparse más que de la distribución de los costes variables.

Por otra parte, c_1 , c_2 y c_3 son los costes unitarios por unidad primaria, por unidad secundaria listada y por unidad secundaria que sea objeto de entrevista o medida, respectivamente.

Como casos particulares típicos de nuestra función de costes tenemos:

1) $a_1 = a_2 = a_3 = 1, \Rightarrow C=c_1n+c_2n\overline{M}+c_3n\overline{m}$

2) Además de verificarse la condición anterior, suponemos $c_2 = 0$, con lo cual no se cuenta el coste del listado de unidades de segunda etapa. Ahora tenemos: $C=c_1n+c_3n\overline{m}$, que suele denominarse *función de coste de campo*, y que es la más utilizada habitualmente.

3) Además de las dos condiciones anteriores suponemos que $c_1 = 0$, lo que equivale a considerar el coste total directamente proporcional al tamaño de la muestra. Tendremos $C=cn\overline{m}=cm$.

Una expresión matemática de la función de coste no deducible de la función general anterior es la *función de coste de Hansen, Hurwitz y Madow*, cuya expresión es $C=c_o\sqrt{n}+c_1n+c_2n\overline{m}$, donde el primer término expresa los gastos de viaje entre las unidades primarias. Hansen, Hurwitz y Madow obtienen el par $\left(n,\overline{m}\right)$ que minimiza la varianza para una función de coste dada.

Nosotros vamos a suponer en los cálculos una función de coste de campo definida como $C = n \cdot c_1 + n \cdot \overline{m} \cdot c_2$, y evaluaremos la varianza de la media a optimizar mediante la expresión aproximada $V(\overline{\overline{x}}) = \frac{S^2}{n\overline{m}}(1 + (\overline{m} - 1) \cdot \delta)$. Para obtener los valores de *n* y $\overline{m}$ que hagan mínima $V(\overline{x})$ con la restricción dada por la función de coste de campo construiremos la función de Lagrange:

$$\phi = \frac{S^2}{n\overline{m}} \cdot (1 + (\overline{m} - 1)\partial) + \lambda\left(C - n \cdot c_1 - n \cdot \overline{m} c_2\right)$$

Igualaremos a cero sus derivadas parciales respecto de *n*, $\overline{m}$ y λ y eliminando parámetros adecuadamente se tiene:

$$\overline{m}_{op} = \sqrt{\frac{c_1}{c_2} \cdot \frac{1 - \delta}{\delta}}$$

MUESTREO BIETÁPICO CON PROBABILIDADES DESIGUALES Y CON REPOSICIÓN EN 1ª ETAPA. ESTIMADORES, VARIANZAS Y SU ESTIMACIÓN

Si consideramos la unidad muestral primaria *i*-ésima de muestreo como una población, siendo $\hat{X}_i$ una estimación de su total al considerar el submuestreo, y representamos por $\overline{x}_i$ un estimador insesgado de su media, podemos aplicar la expresión del estimador general de Hansen y Hurwitz $\hat{X}_{HH}$ (estudiado en el Capítulo 2) al muestreo bietápico, siendo la primera etapa con reposición (la segunda etapa puede ser con o sin reposición). Así, un estimador insesgado del total será:

Un estimador insesgado del total será: $\hat{\hat{X}}_{HH} = \sum_i^n \frac{\hat{X}_i}{nP_i} = \frac{1}{n}\sum_i^n \frac{\hat{X}_i}{P_i} = \frac{1}{n}\sum_i^n \frac{M_i \overline{x}_i}{P_i}$.

Para probabilidades proporcionales al tamaño $\rightarrow P_i = \frac{M_i}{M}$ con $M = \sum_{i=1}^{N} M_i$, luego:

$$\hat{\hat{X}}_{HH} = \frac{1}{n}\sum_i^n \frac{M_i \overline{x}_i}{P_i} = \frac{1}{n}\sum_i^n \frac{M_i \overline{x}_i}{M_i / M} = \frac{M}{n}\sum_i^n \overline{x}_i$$

Los *estimadores para medias, proporciones y totales de clase en el muestreo bietápico con probabilidades desiguales* son inmediatos:

$$\hat{\hat{\overline{X}}} = \frac{1}{M}\hat{\hat{X}}_{HH} = \frac{1}{M}\sum_i^n \frac{\hat{X}_i}{nP_i} = \frac{1}{n}\sum_i^n \frac{\frac{M_i}{M}\overline{x}_i}{P_i}, \quad \hat{\hat{A}} = M\hat{\hat{P}} = M\frac{1}{n}\sum_i^n \frac{\frac{M_i}{M}\hat{P}_i}{P_i} = \frac{1}{n}\sum_i^n \frac{M_i \hat{P}_i}{P_i}$$

$$\hat{\hat{P}} = \frac{1}{n}\sum_i^n \frac{\frac{M_i}{M}\hat{P}_i}{P_i}$$

$\hat{P}_i$ = proporción muestral en el conglomerado *i*-ésimo.

Varianzas

Como la primera etapa es siempre con reposición, distinguiremos entre si la segunda etapa es con reposición o sin reposición.

Sin reposición en segunda etapa

$$V(\hat{\hat{X}}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{X_i}{P_i} - X\right)^2 P_i + \sum_{i}^{N}\frac{M_i^2(1-f_{2i})}{nP_i m_i}\cdot S_i^2, \quad V(\hat{\hat{\bar{X}}}_{HH}) = \frac{1}{M^2}V(\hat{\hat{X}}_{HH})$$

$$V(\hat{\hat{A}}_{HH}) = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{A_i}{P_{ri}} - A^2\right) + \sum_{i}^{N}\frac{M_i^2(1-f_{2i})}{nP_{ri}m_i}\cdot\frac{M_i P_i Q_i}{M_i - 1}$$

$$V(\hat{\hat{P}}_{HH}) = \frac{1}{M^2}V(\hat{\hat{A}})$$

Para el caso particular de ***probabilidades proporcionales a los tamaños*** $P_i = \dfrac{M_i}{M}$ con $M = \sum_{i=1}^{N} M_i$, se tiene:

$$V(\hat{\hat{X}}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{X_i}{M_i/M} - X\right)^2\frac{M_i}{M} + \sum_{i}^{N}\frac{M_i^2(1-f_{2i})}{nm_i M_i/M}\cdot S_i^2 = \frac{M}{n}\left[\sum_{i=1}^{N}\left(\frac{X_i^2}{M_i} - \frac{X^2}{M}\right)^2 + \sum_{i}^{N}\frac{M_i}{m_i}(1-f_{2i})\cdot S_i^2\right]$$

Con reposición en segunda etapa

$$V(\hat{\hat{X}}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{X_i}{P_i} - X\right)^2 P_i + \sum_{i}^{N}\frac{M_i^2}{nP_i m_i}\cdot\sigma_i^2, \quad V(\hat{\hat{\bar{X}}}_{HH}) = \frac{1}{M^2}V(\hat{\hat{X}}_{HH})$$

$$V(\hat{\hat{A}}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{A_i}{P_{ri}} - A\right)^2 P_{ri} + \sum_{i}^{N}\frac{M_i^2}{nP_{ri}m_i}\cdot P_i Q_i$$

$$V(\hat{\hat{P}}_{HH}) = \frac{1}{M^2}V(\hat{\hat{A}})$$

Para el caso particular de ***probabilidades proporcionales a los tamaños*** $P_i = \dfrac{M_i}{M}$ con $M = \sum_{i=1}^{N} M_i$, se tiene:

$$V(\hat{\hat{X}}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{X_i}{M_i/M} - X\right)^2\frac{M_i}{M} + \sum_{i}^{N}\frac{M_i^2}{nm_i M_i/M}\cdot\sigma_i^2 = \frac{M}{n}\left[\sum_{i=1}^{N}\left(\frac{X_i^2}{M_i} - \frac{X^2}{M}\right)^2 + \sum_{i}^{N}\frac{M_i}{m_i}\cdot\sigma_i^2\right]$$

Estimación de varianzas (obtenidas por el método de los conglomerados últimos)

Los estimadores insesgados para las varianzas de los estimadores cuando la primera etapa es con reposición, no dependen de si la segunda etapa es o no con reposición.

Independientemente de que la segunda etapa sea o no con reposición, si la primera etapa es con reposición, los estimadores insesgados para las varianzas de los estimadores son los siguientes:

$$\hat{V}\left(\hat{\hat{X}}\right)=\frac{\sum_{i}^{n}\left(\frac{\hat{X}_i}{P_i}-\hat{\hat{X}}_{HH}\right)^2}{n(n-1)},\ \hat{V}(\hat{\bar{X}})=\frac{1}{M^2}\hat{V}(\hat{\hat{X}}_{HH})$$

$$\hat{V}\left(\hat{\hat{A}}\right)=\frac{\sum_{i}^{n}\left(\frac{\hat{A}_i}{P_i}-\hat{\hat{A}}\right)^2}{n(n-1)}=\frac{\sum_{i}^{n}\left(\frac{M_i\hat{P}_i}{P_i}-M\hat{\hat{P}}\right)^2}{n(n-1)}$$

$$\hat{V}(\hat{\hat{P}})=\frac{1}{M^2}\hat{V}(\hat{\hat{A}})$$

MUESTREO BIETÁPICO CON PROBABILIDADES DESIGUALES Y SIN REPOSICIÓN EN 1ª ETAPA. ESTIMADORES, VARIANZAS Y SU ESTIMACIÓN

Si consideramos la unidad muestral primaria *i*-ésima de muestreo como una población, siendo $\hat{X}_i$ una estimación de su total al considerar el submuestreo, y representamos por $\bar{x}_i$ un estimador insesgado de su media, podemos aplicar la expresión del estimador general de Hoewitz y Thompson $\hat{X}_{HT}$ al muestreo bietápico, siendo la primera etapa sin reposición (la segunda etapa puede ser con o sin reposición). Así, un estimador insesgado del total será:

$$\hat{\hat{X}}_{HT}=\sum_{i}^{n}\frac{\hat{X}_i}{\pi_i}=\sum_{i}^{n}\frac{M_i\bar{x}_i}{\pi_i}$$

Como casos particulares de este estimador tenemos:

Conglomerados del mismo tamaño $\overline{M}$

$$\hat{\hat{X}}_{HT}=\sum_{i}^{n}\frac{\overline{M}\bar{x}_i}{\pi_i}=\overline{M}\sum_{i}^{n}\frac{\bar{x}_i}{\pi_i}$$

Probabilidades proporcionales al tamaño $\rightarrow \pi_i=\frac{nM_i}{M}$ con $M=\sum_{i=1}^{N}M_i$

$$\hat{\hat{X}}_{HT}=\sum_{i}^{n}\frac{M_i\bar{x}_i}{\pi_i}=\sum_{i}^{n}\frac{M_i\bar{x}_i}{nM_i/M}=\frac{M}{n}\sum_{i}^{n}\bar{x}_i$$

Probabilidades iguales → $\pi_i = \dfrac{n}{N}$

$$\hat{\bar{X}}_{HT} = \sum_i^n \frac{M_i \bar{x}_i}{\pi_i} = \sum_i^n \frac{M_i \bar{x}_i}{n/N} = \frac{N}{n} \sum_i^n M_i \bar{x}_i$$

Vemos que las expresiones de los estimadores coinciden en muestreo con y sin reposición.

Los *estimadores para medias, proporciones y totales de clase en el muestreo bietápico con probabilidades desiguales* son inmediatos:

$$\hat{\bar{\bar{X}}} = \frac{1}{M} \hat{\bar{X}}_{HT} = \frac{1}{M} \sum_i^n \frac{\hat{X}_i}{\pi_i} = \sum_i^n \frac{\dfrac{M_i}{M} \bar{x}_i}{\pi_i}$$

$$\hat{\bar{\bar{P}}} = \sum_i^n \frac{\dfrac{M_i}{M} \hat{P}_i}{\pi_i} \quad \hat{P}_i = \text{proporción muestral en el conglomerado } i\text{-ésimo}$$

$$\hat{\bar{\bar{A}}} = M\hat{\bar{\bar{P}}} = M \sum_i^n \frac{\dfrac{M_i}{M} \hat{P}_i}{\pi_i} = \sum_i^n \frac{M_i \hat{P}_i}{\pi_i}$$

Varianzas

Como la primera etapa es siempre sin reposición, distinguiremos entre si la segunda etapa es con reposición o sin reposición.

Sin reposición en segunda etapa

$$V(\hat{\bar{\bar{X}}}_{HT}) = \sum_{i=1}^N \frac{X_i^2}{\pi_i}(1-\pi_i) + \sum_{i \neq j}^N \frac{X_i}{\pi_i} \frac{X_j}{\pi_j}(\pi_{ij} - \pi_i \pi_j) + \sum_i^N \frac{(1-f_{2i}) M_i^2 S_i^2}{m_i \pi_i},$$

$$V(\hat{\bar{\bar{X}}}_{HH}) = \frac{1}{M^2} V(\hat{\bar{X}}_{HH})$$

Con reposición en segunda etapa

$$V(\hat{\bar{\bar{X}}}_{HT}) = \sum_{i=1}^N \frac{X_i^2}{\pi_i}(1-\pi_i) + \sum_{i \neq j}^N \frac{X_i}{\pi_i} \frac{X_j}{\pi_j}(\pi_{ij} - \pi_i \pi_j) + \sum_i^N \frac{M_i^2 \sigma_i^2}{m_i \pi_i}, \quad V(\hat{\bar{\bar{X}}}_{HH}) = \frac{1}{M^2} V(\hat{\bar{\bar{X}}}_{HH})$$

Para el caso particular de totales de clase y proporciones se hacen las siguintes sustituciones en las fórmulas anteriores:

$$S_i^2 = \frac{M_i}{M_i - 1} P_i Q_i, \quad \sigma_i^2 = P_i Q_i$$

Estimación de varianzas (obtenidas mediante los teoremas I y II de Durbin)

Sin reposición en segunda etapa

$$\hat{V}(\hat{\hat{X}}_{HT}) = \sum_{i=1}^{n} \frac{\hat{X}_i^2}{\pi_i}(1-\pi_i) + \sum_{i\neq j}^{n} \frac{\hat{X}_i}{\pi_i}\frac{\hat{X}_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j) + \sum_{i}^{n} \frac{(1-f_{2i})M_i^2\hat{S}_i^2}{m_i\pi_i}$$

Con reposición en segunda etapa

$$\hat{V}(\hat{\hat{X}}_{HT}) = \sum_{i=1}^{N} \frac{\hat{X}_i^2}{\pi_i}(1-\pi_i) + \sum_{i\neq j}^{N} \frac{\hat{X}_i}{\pi_i}\frac{\hat{X}_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j) + \sum_{i}^{N} \frac{M_i^2\hat{S}_i^2}{m_i\pi_i}$$

Para las medias se hace $\hat{V}(\hat{\bar{X}}_{HH}) = \frac{1}{M^2}\hat{V}(\hat{\hat{X}}_{HH})$.

Para el caso particular de totales de clase y proporciones se hace $\hat{S}_i^2 = \frac{m_i}{m_i-1}\hat{P}_i\hat{Q}_i$.

MUESTREO POLIETÁPICO

En el muestreo polietápico se realizan submuestreos consecutivos hasta un número de etapas determinado. Por ejemplo, en el muestreo trietápico se selecciona en una primera etapa una muestra de unidades primarias, en una segunda etapa se realiza submuestreo en cada una de las unidades de la muestra de primera etapa y en una tercera etapa se realiza submuestreo en cada una de las unidades de la muestra de segunda etapa. De forma similar se generalizaría para un número elevado de etapas, dando lugar al muestreo polietápico.

Muestreo con reposición de unidades primarias y sin reposición en las restantes etapas

Considerando la unidad muestral *i*-ésima como una población y representando por $\bar{x}_i$ un estimador insesgado de $\bar{X}_i$, podemos extender el estimador insesgado de Hansen y Hurwitz a cualquier número de etapas. Tenemos entonces que un estimador insesgado del total será:

$$\hat{\hat{X}}_{HH} = \sum_{i}^{n} \frac{\hat{X}_i}{nP_i} = \frac{1}{n}\sum_{i}^{n} \frac{\hat{X}_i}{P_i} = \frac{1}{n}\sum_{i}^{n} \frac{M_i\bar{x}_i}{P_i}$$

La varianza de este estimador y su estimación son las siguientes:

$$V(\hat{\hat{X}}_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{X_i}{P_i} - X\right)^2 P_i + \sum_{i}^{N} nP_i\cdot\sigma_i^2 \qquad \hat{V}\left(\hat{\hat{X}}\right) = \frac{\sum_{i}^{n}\left(\frac{\hat{X}_i}{P_i} - \hat{\hat{X}}_{HH}\right)^2}{n(n-1)}$$

Muestreo sin reposición en todas las etapas

Considerando la unidad muestral *i*-ésima como una población y representando por $\hat{X}_i$ un estimador insesgado de X_i, podemos extender el estimador insesgado de Horvitz y Thompson a cualquier número de etapas. Tenemos entonces que un estimador insesgado del total será:

$$\hat{\hat{X}}_{HT} = \sum_i^n \frac{\hat{X}_i}{\pi_i} = \sum_i^n \frac{M_i \bar{x}_i}{\pi_i} = \sum_i^n \frac{M_i \bar{x}_i}{n/N} = \frac{N}{n} \sum_i^n M_i \bar{x}_i$$

La varianza de este estimador es:

$$V(\hat{\hat{X}}_{HT}) = \sum_{i=1}^N \frac{X_i^2}{\pi_i^2} \pi_i + 2 \sum_{i<j}^N \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} \pi_{ij} - X^2 + \sum_i^N \sigma_i^2 \pi_i$$

Un estimador insesgado para la varianza es:

$$\hat{V}(\hat{\hat{X}}_{HT}) = \sum_{i=1}^n \frac{\hat{X}_i^2}{\pi_i}(1-\pi_i) + \sum_{i \neq j}^n \frac{\hat{X}_i}{\pi_i} \frac{\hat{X}_j}{\pi_j}(\pi_{ij} - \pi_i \pi_j) + \sum_i^n \frac{(1-f_{2i}) M_i^2 \hat{S}_i^2}{m_i \pi_i}$$

DISEÑOS COMPLEJOS: MUESTREO BIETÁPICO CON ESTRATIFICACIÓN EN PRIMERA ETAPA

En la práctica es habitual utilizar diseños polietápicos con distintos tipos de muestreo en cada etapa. Es muy común utilizar estratificación de unidades primarias para seleccionar las unidades primarias de la muestra de primera etapa mediante muestreo estratificado. Después se realiza la selección de las unidades de segunda etapa dentro de cada unidad de primera etapa. Para este tipo de muestreo bietápico con estratificación en primera etapa las fórmulas de los estimadores, varianzas y estimaciones de varianzas se presentarán a continuación.

Sean los pesos de los estratos y las fracciones de muestreo.

$$W_h = \frac{N_h \overline{M}_h}{N \overline{M}} f_h = \frac{n_h \overline{m}_h}{N_h \overline{M}_h} = f_{1h} \cdot f_{2h}$$

Un estimador insesgado de la media es $\bar{\bar{x}}_{st} = \sum_h^L W_h \bar{\bar{x}}_h = \sum_h^L W_h \cdot \frac{1}{n_h} \sum_i^{n_h} \bar{x}_{ih}$ pues

$$E(\bar{\bar{x}}_{st}) = \sum_h^L W_h E_1 E_2 \bar{\bar{x}}_h = \sum_h^L W_h E_1 \frac{1}{n_h} \sum_i^n E_2 \bar{x}_{ih} = \sum_h^L W_h E_1 \bar{x}_h = \sum_h^L W_h \overline{X}_h = \bar{\bar{X}}$$

La varianza del estimador de la media viene dada por:

$$V(\bar{\bar{x}}_{st}) = \sum_h^L W_h^2 \cdot V(\bar{x}_h) = \sum_i^L W_h^2 \left[(1-f_{1h}) \cdot \frac{S_{bh}^2}{n_h \overline{M}_h} + (1-f_{2h}) \cdot \frac{S_{wh}^2}{n_h \overline{m}_h} \right]$$

La muestra es autoponderada si $f_h = f_{1h} \cdot f_{2h} = f$ y la estimación de la varianza vendrá dada por la siguiente expresión:

$$\hat{V}(\bar{\bar{x}}_{st}) = \sum_h^L W_h^2 \cdot \hat{V}(\bar{x}_h) = \sum_i^L W_h^2 \left[(1-f_{1h}) \cdot \frac{\hat{S}_{bh}^2}{n_h \overline{m}_h} + f_{1h}(1-f_{2h}) \cdot \frac{S_{wh}^2}{n_h \overline{m}_h} \right]$$

De forma similar se realizan otros diseños complejos de encuestas. En cada etapa se aplicarán los cálculos relativos al tipo de muestreo definido en ella.

PROBLEMAS RESUELTOS

8.1. En un barrio de una ciudad se obtiene una muestra de 6 manzanas de 30 casas cada una con probabilidades iguales. Dentro de cada manzana de la muestra se realiza submuestreo sin reposición con fracción de muestreo igual a 1/6, y se obtienen los siguientes valores para el número de casas en las que viven jubilados:

Manzana	1	2	3	4	5	6
Nº de casas con jubilados	4	3	5	2	1	5

Se pide:

1) Suponiendo muestreo con reposición de unidades primarias, estimar la proporción *P* de casas del barrio en las que viven jubilados y su error relativo de muestreo. Estimar por intervalos al 95% el total *A* de casas del barrio en las que viven jubilados.
2) Suponiendo muestreo sin reposición de unidades primarias y fracción de muestreo en primera etapa igual a 1/2, estimar la proporción de casas del barrio en las que viven jubilados y su error relativo de muestreo. Construir la tabla del análisis de la varianza para la muestra y estimar el valor del coeficiente de correlación intraconglomerados. Estimar por intervalos al 95% el total de casas del barrio en las que viven jubilados.

Consideramos las manzanas como conglomerados de igual tamaño (30 casas cada manzana).

Tenemos como datos $n = 6,\ \overline{M} = 30,\ f_{2i} = \dfrac{m_i}{\overline{M}} \Rightarrow m_i = f_{2i}\,\overline{M} = \dfrac{1}{6}30 = 5 = \overline{m}$.

Estamos entonces en muestreo bietápico de conglomerados del mismo tamaño con submuestreo también del mismo tamaño y con reposición en primera etapa sin existir reposición en segunda etapa. El estimador de la proporción es:

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n}\hat{P}_i = \frac{1}{6}\left(\frac{4}{5}+\frac{3}{5}+\frac{5}{5}+\frac{2}{5}+\frac{1}{5}+\frac{5}{5}\right) = \frac{2}{3}$$

Para calcular la varianza del estimador realizamos la tabla muestral del análisis de la varianza. Para ello utilizamos seis variables de clasificación de, *C*1 a *C*6*,* una por cada conglomerado muestral, de modo que cada variable tiene un número de unos igual al total de clase del conglomerado muestral correspondiente, y ceros para el resto de las unidades del conglomerado muestral. Se elige *Análisis de la varianza de un factor* en *Análisis de datos* del menú *Herramientas,* y se rellena su pantalla de entrada como se indica en la Figura 8-1. Los resultados se ven en la Figura 8-2.

La varianza es $\hat{V}(\hat{P}) = \dfrac{\hat{S}_b^2}{n\overline{m}} = \dfrac{0{,}53333}{6{,}5} = 0{,}018$. El error relativo de muestreo es:

$$Cv(\hat{P}) = \frac{\sqrt{\hat{V}(\hat{P})}}{\hat{P}} = \frac{\sqrt{0{,}018}}{2/3} = \frac{0{,}134164}{2/3} = 0{,}2 \quad (20\%)$$

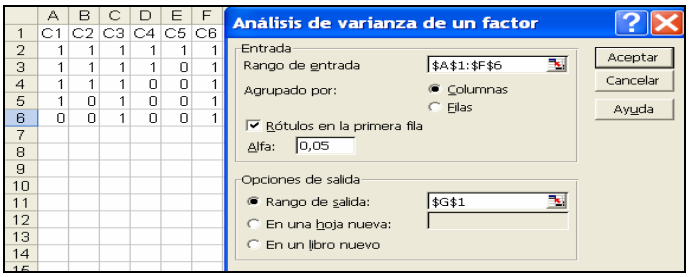


Figura 8-1

	G	H	I	J	K	L	M
1	Análisis de varianza de un factor						
2							
3	RESUMEN						
4	*Grupos*	*Cuenta*	*Suma*	*Promedio*	*Varianza*		
5	C1	5	4	0,8	0,2		
6	C2	5	3	0,6	0,3		
7	C3	5	5	1	0		
8	C4	5	2	0,4	0,3		
9	C5	5	1	0,2	0,2		
10	C6	5	5	1	0		
11							
12							
13	ANÁLISIS DE VARIANZA						
14	*Origen de las variaciones*	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
15	Entre grupos	2,666666667	5	0,533333333	3,2	0,023804484	2,62065214
16	Dentro de los grupos	4	24	0,166666667			
17							
18	Total	6,666666667	29				
19							

Figura 8-2

Al ser la fracción de muestreo en primera etapa 1/2, tenemos $1/2 = 6/N$, de donde el número de conglomerados en la población es $N = 13$. Para hacer una estimación por intervalos del total de la característica A en la población, necesitamos la varianza del estimador del total. Pero:

$$\hat{V}(\hat{A}) = N^2 \overline{M}^2 \hat{V}(\hat{P}) = 12^2 * 30^2 * 0{,}018 = 2332{,}8 \Rightarrow \hat{\sigma}(\hat{A}) = 48{,}3$$

El intervalo de confianza para el total al 95% suponiendo normalidad será:

$$\left(\hat{A} - \lambda_\alpha \hat{\sigma}(\hat{A}),\ \hat{A} - \lambda_\alpha \hat{\sigma}(\hat{A})\right) = \left[240 - 1{,}96 * 48{,}3,\ \ 240 + 1{,}96 * 48{,}3\right] = \left[145{,}33,\ \ 334{,}66\right]$$

$$\hat{A} = N\overline{M}\hat{P} = 12 * 30 * \frac{2}{3} = 240$$

En el caso de que ambas etapas sean sin reposición, los estimadores de la proporción y el total de clase no varían, pero sí cambian los errores de muestreo. La varianza del estimador de la proporción será ahora:

$$\hat{V}\left(\hat{P}\right) = \left(1 - f_1\right)\frac{\hat{S}_b^2}{r\overline{m}} + f_1\left(1 - f_2\right)\frac{\hat{S}_w^2}{r\overline{m}} = \left(1 - \frac{1}{2}\right)\frac{0{,}5333}{6{,}5} + \frac{1}{2}\left(1 - \frac{1}{6}\right)\frac{0{,}1666}{6{,}5} = 0{,}0112$$

El error relativo es $Cv(\hat{P}) = \dfrac{\sqrt{\hat{V}(\hat{P})}}{\hat{P}} = \dfrac{\sqrt{0{,}0112}}{2/3} = \dfrac{0{,}10583}{2/3} = 0{,}1587$ (15,87%) y se observa que en muestreo sin reposición el error resulta ser menor.

8.2. Una región tiene 1000 hogares agrupados en 50 pequeños municipios de tamaños desiguales M_i (i = 1, 2, ..., 50). Se trata de estimar la proporción de hogares que están al corriente de sus obligaciones fiscales mediante muestreo de conglomerados con submuestreo con probabilidades iguales y sin reposición en las dos etapas. En la primera etapa se obtienen 5 municipios muestrales de tamaños 6, 10, 8, 20 y 60 hogares. En la segunda etapa, realizada con fracciones de muestreo $f_{2i} = 4/M_i$, se obtiene en los 5 municipios de la muestra de primera etapa los valores 1, 3, 2, 2 y 3 para el número de hogares que están al corriente de sus obligaciones fiscales. Se pide:

1) Hallar el estimador insesgado de la proporción de hogares que están al corriente de sus obligaciones fiscales y su error absoluto y relativo de muestreo.

2) Construir la tabla del análisis de la varianza para la muestra y comprobar la igualdad fundamental.

Consideramos los municipios como conglomerados de distinto tamaño. Las unidades elementales son los hogares de los municipios. Tenemos:

$$f_{2i} = \frac{m_i}{M_i} = \frac{4}{M_i} \Rightarrow m_i = 4 \,\forall i$$

El estimador insesgado para la proporción en muestreo bietápico para conglomerados de distinto tamaño es:

$$\hat{P} = \frac{N}{n}\sum_i^n \frac{M_i}{M}\hat{P}_i = \frac{50}{5}\cdot\frac{1}{1000}\sum_i^5 M_i\hat{P}_i = \frac{1}{100}\left(6\frac{1}{4}+10\frac{3}{4}+8\frac{2}{4}+20\frac{2}{4}+60\frac{3}{4}\right) = 0{,}68$$

Para estimar la varianza de la proporción utilizamos la fórmula adecuada al muestreo bietápico sin reposición en las dos etapas con probabilidades iguales para conglomerados de distinto tamaño. Tenemos:

$$\hat{V}\left(\hat{P}\right) = \frac{1}{M^2}\left[\frac{N^2(1-f_1)}{n}\cdot\frac{\sum_i^n\left(M_i\hat{P}_i - \frac{1}{n}\sum_{i=1}^n M_i\hat{P}_i\right)^2}{n-1} + \frac{N}{n}\sum_i^n M_i^2(1-f_{2i})\cdot\frac{P_iQ_i}{m_i-1}\right] = 0{,}1458$$

El error relativo de muestreo viene dado por el coeficiente de variación del estimador. Tenemos:

$$Cv(\hat{P}) = \frac{\sqrt{\hat{V}(\hat{P})}}{\hat{P}} = \frac{\sqrt{0{,}1458}}{0{,}68} = \frac{0{,}38}{0{,}68} = 0{,}5588\ (55{,}88\%)$$

Como $m_i = 4 = \overline{m} \;\; \forall i$, la tabla del análisis de la varianza para la muestra en este caso del muestreo bietápico es la siguiente:

Fuente	Grados libertad	Sumas de cuadrados	Cuadrados medios
"entre"	$n-1$	$\overline{m}\sum_{i}^{n}\left(\hat{P}_i-\overline{P}\right)^2$	$\hat{S}_b^2$
"dentro"	$n(\overline{m}-1)$	$\overline{m}\sum_{i=1}^{n}\hat{P}_i\left(1-\hat{P}_i\right)$	$\hat{S}_w^2$
Total	$n\overline{m}-1$	$n\overline{m}\overline{P}\overline{Q}$	$\hat{S}^2$

La relación fundamental del análisis de la varianza será: $(n\overline{m}-1)\hat{S}^2=(n\overline{m}-n)\hat{S}_w^2+(n-1)\hat{S}_b^2$. Todos los elementos del cuadro son calculables con nuestros datos, con lo que ya pueden realizarse las operaciones para obtener los siguientes resultados:

Fuente	Grados libertad	Sumas de cuadrados	Cuadrados medios
"entre"	$5-1=4$	0,7	0,175
"dentro"	$5(4-1)=15$	4,25	0,2833
Total	$5\cdot 4-1=19$	4,95	0,26

8.3. Consideremos una provincia con 400 municipios. Para estimar el total de hogares con automóvil en la provincia se selecciona una muestra de 10 municipios con igual probabilidad, y dentro de cada municipio de la muestra se seleccionan aleatoriamente hogares utilizando una fracción de muestreo $f = 1/5$. Se obtienen los siguientes datos:

Distritos muestrales	*Total de hogares en los distritos* (M_i)	*Nº de hogares en la muestra* (m_i)	*Hogares con coche* (A_i)
1	200	40	6
2	180	35	7
3	35	7	1
4	220	44	7
5	80	16	1
6	140	28	3
7	125	25	2
8	65	13	2
9	140	28	2
10	55	11	1

Se pide:

1) Estimar el total de hogares con automóvil en la provincia y sus errores absoluto y relativo de muestreo.

2) Realizar la estimación anterior por intervalos al 95% de confianza.

Consideramos los municipios como conglomerados de distinto tamaño. Las unidades elementales son los hogares dentro de los municipios.

El estimador insesgado para la proporción en muestreo bietápico para conglomerados de distinto tamaño con probabilidades iguales es:

$$\hat{A} = \frac{N}{n}\sum_{i}^{n} M_i \hat{P}_i = \frac{400}{10}\left(200\frac{6}{40} + 180\frac{7}{35} + \cdots + 53\frac{1}{11}\right) = 6440$$

Para estimar la varianza del total de clase utilizamos la fórmula adecuada al muestreo bietápico sin reposición en las dos etapas (no se especifica otra cosa) con probabilidades iguales para conglomerados de distinto tamaño. Tenemos:

$$\hat{V}\left(\hat{A}\right) = \frac{N^2(1-f_1)}{n}\cdot\frac{\sum_{i}^{n}\left(M_i\hat{P}_i - \frac{1}{n}\sum_{i=1}^{n} M_i\hat{P}_i\right)^2}{n-1} + \frac{N}{n}\sum_{i}^{n} M_i^2(1-f_{2i})\cdot\frac{P_iQ_i}{m_i - 1} = 628237$$

El error relativo de muestreo viene dado por el coeficiente de variación del estimador. Tenemos:

$$Cv(\hat{P}) = \frac{\sqrt{\hat{V}(\hat{A})}}{\hat{A}} = \frac{\sqrt{628237}}{6440} = \frac{792{,}614}{6440} = 0{,}123 \quad (12{,}3\%)$$

Para hacer una estimación por intervalos del total de la característica suponiendo normalidad tendremos:

$$\left(\hat{A} - \lambda_\alpha\hat{\sigma}(\hat{A}), \hat{A} - \lambda_\alpha\hat{\sigma}(\hat{A})\right) = \left[6440 - 1.96\cdot 792{,}61,\ 6440 + 1.96\cdot 792{,}61\right] = \left[4886.4,\ 7993.5\right]$$

8.4. De una viña formada por 1000 líneos de 50 cepas cada uno, se extrae una muestra de 30 líneos. Dentro de cada líneo de la muestra se analizan cinco cepas, utilizando muestreo con probabilidades iguales y con reemplazamiento en primera etapa. El análisis de la varianza de la muestra para una variable medida sobre las cepas presenta los siguientes resultados:

Fuente de variación	Grados de libertad	Cuadrados medios
Entre líneos	29	600
Dentro de líneos	120	400

1) Estimar el error de muestreo del estimador de la media de la variable medida sobre las cepas. Hallar la amplitud de las estimaciones por intervalos al 95% de confianza.

2) Realizar los mismos cálculos para muestreo sin reposición en ambas etapas, comparando los resultados con los del apartado anterior.

Consideramos cada líneo como conglomerado de 50 cepas (tamaños iguales). Cuando existe reposición en primera etapa, la fórmula de la estimación de la varianza de la media, independientemente de que haya o no reposición en segunda etapa, es la siguiente:

$$\hat{V}\left(\bar{\bar{x}}\right) = \frac{\hat{S}_b^2}{n\overline{m}}$$

La tabla del análisis de la varianza para la muestra en el caso del muestreo bietápico es la siguiente:

Fuente	Grados libertad	Sumas de cuadrados	Cuadrados medios
"entre"	$n-1$	$\overline{m}\sum_{i}^{n}\left(\bar{x}_i-\bar{\bar{x}}\right)^2$	$\hat{S}_b^2$
"dentro"	$n(\overline{m}-1)$	$\sum_{i}^{n}\sum_{j}^{\overline{m}}\left(X_{ij}-\bar{x}_i\right)^2$	$\hat{S}_w^2$
Total	$n\overline{m}-1$	$\sum_{i}^{n}\sum_{j}^{\overline{m}}\left(X_{ij}-\bar{\bar{x}}\right)^2$	$\hat{S}^2$

Si consideramos los datos de nuestro problema tenemos $\hat{S}_b^2 = 600$ y $\hat{S}_w^2 = 400$. Por tanto:

$$\hat{V}\left(\bar{\bar{x}}\right)=\frac{\hat{S}_b^2}{n\overline{m}}=\frac{600}{29\cdot 5}=4$$

La amplitud del intervalo de confianza al 95% es $2\sqrt{\hat{V}\left(\bar{\bar{x}}\right)}$, que puede considerarse como un límite para el error de muestreo, y que en nuestro caso vale 4.

Si las dos etapas son sin reposición se tiene:

$$\hat{V}\left(\bar{\bar{x}}\right)=\left(1-f_1\right)\frac{\hat{S}_b^2}{n\overline{m}}+f_1\left(1-f_2\right)\cdot\frac{\hat{S}_w^2}{n\overline{m}}=\left(1-\frac{30}{1000}\right)\frac{600}{30\cdot 5}+\frac{30}{1000}\left(1-\frac{5}{50}\right)\cdot\frac{400}{30\cdot 5}=3{,}95$$

La amplitud del intervalo de confianza al 95% es $2\sqrt{\hat{V}\left(\bar{\bar{x}}\right)}$, que en este caso vale 7,9. Como es natural, tiene menos varianza el muestreo sin reposición, ya que siempre es más preciso. Este hecho también se refleja en la anchura de los intervalos de confianza.

8.5. Un fabricante de prendas de vestir tiene 90 plantas localizadas en todo Estados Unidos y quiere estimar el número promedio de horas que las máquinas de coser estuvieron sin funcionar por reparación en los meses pasados. Debido a que las plantas están muy dispersas, el fabricante decide utilizar un muestreo por conglomerados, especificando cada planta como un conglomerado de máquinas. Cada planta contiene muchas máquinas, y el verificar los registros de reparación de cada máquina implicaría consumir tiempo. Por tanto el fabricante usa un muestreo en dos etapas. Se dispone de tiempo y dinero suficientes para muestrear 10 plantas y aproximadamente un 20% de las máquinas de cada planta. Dados los siguientes datos sobre el tiempo sin funcionar para las máquinas de coser por plantas

Planta	M_i	m_i	Tiempo sin funcionar (en horas)	$\bar{x}_i$	S^2_i
1	so	10	5, 7, 9, 0, 11, 2, 8, 4, 3, 5	5,40	11,38
2	65	13	4, 3, 7, 2, 11, 0, 1, 9, 4, 3, 2, 1, 5	4,00	10,67
3	45	9	5, 6, 4, 11, 12, 0, 1, 8, 4	5,67	16,75
4	48	10	6, 4, 0, 1, 0, 9, 8, 4, 6, 10	4,80	13,29
5	52	10	11, 4, 3, 1, 0, 2, 8, 6, 5, 3	4,30	11,12
6	58	12	12, 11, 3, 4, 2, 0, 0, 1, 4, 3, 2, 4	3,83	14,88
7	42	8	3, 7, 6, 7, 8, 4, 3, 2	5,00	5,14
8	66	13	3, 6, 4, 3, 2, 2, 8, 4, 0, 4, 5, 6, 3	3,85	4,31
9	40	8	6, 4, 7, 3, 9, 1, 4, 5	4,88	6,13
10	56	11	6, 7, 5, 10, 11, 2, 1, 4, 0, 5, 4	5,00	11,80

Estimar el tiempo sin funcionar promedio por máquina y establecer un límite para el error de estimación. El fabricante sabe que tiene un total de 4.500 máquinas en todas las plantas. Estimar también la cantidad total de tiempo sin funcionar durante el mes pasado para todas las máquinas. Estimar el tiempo sin funcionar promedio por máquina en caso de que no se conozca el número total de máquinas.

Para estimar el tiempo promedio sin funcionar por máquina tenemos:

$$\bar{\bar{x}} = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{M}\bar{x}_i = \frac{90}{4500\cdot 10}(50\cdot 5{,}4 + 65\cdot 4 + \cdots + 56\cdot 5) = 4{,}8$$

$$\hat{V}(\bar{\bar{x}}) = \frac{N^2(1-f_1)}{nM^2}\cdot\frac{\sum_i^n\left(\hat{X}_i - \hat{\bar{X}}_i\right)^2}{n-1} + \frac{N}{nM^2}\sum_i^n\frac{M_i^2(1-f_{2i})}{m_i}\cdot\frac{\sum_j^{m_i}\left(X_{ij}-\bar{x}_i\right)^2}{m_i-1} =$$

$$\frac{90^2\left(1-\frac{10}{90}\right)}{10\cdot 4500^2}\cdot 768{,}38 + \frac{90}{10\cdot 4500^2}\cdot 21990{,}96 = 0{,}037094$$

Un límite para el error de estimación puede calcularse a través del intervalo de confianza para el estimador $\bar{\bar{x}} \pm 2\sqrt{0{,}037094} = 4{,}8 \pm 0{,}38$.

Para la estimación de la cantidad total de tiempo sin funcionar para todas las máquinas tenemos el estimador $\hat{X} = M\bar{\bar{x}} = 4500\cdot 4{,}8 = 21600$, siendo la estimación de su varianza $\hat{V}(\hat{X}) = M^2V(\bar{\bar{x}}) = 4500^2\cdot 0{,}037094 = 751153{,}5$.

Si no se conoce M se estima la media mediante el estimador de razón:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{n}M_i\bar{x}_i}{\sum_{i=1}^{n}M_i} = \frac{(50\cdot 5{,}4 + 65\cdot 4 + \cdots + 56\cdot 5)}{50 + 65 + \cdots 56} = 4{,}6$$

$$\hat{V}(\bar{\bar{x}}) = \frac{1-f}{n\bar{M}^2}(\hat{S}_x^2 + \hat{R}^2S_M^2 - 2\hat{R}\hat{S}_{xm}) = \frac{1-f}{n\bar{M}^2(n-1)}\left(\sum_{i=1}^{10}\left(M_i\bar{x}_i\right)^2 + \bar{\bar{x}}^2\sum_{i=1}^{10}M_i^2 - 2\bar{\bar{x}}\sum_{i=1}^{10}M_i\bar{\bar{x}}_i\,M_i\right) = 0{,}049$$

Se observa que la estimación por razón, provocada por el desconocimiento de M, origina un error superior, pero no en demasiada cuantía.

8.6. Para estimar el total de una magnitud en una población de 100 conglomerados se estratifica la misma en dos zonas, rural y urbana, con 60 y 40 conglomerados respectivamente. En la zona rural se selecciona una muestra de cinco conglomerados con probabilidades proporcionales a su tamaño M_i y con reemplazamiento, mientras que en la zona urbana se selecciona una muestra sistemática de cuatro conglomerados con coeficiente de correlación intramuestral igual a una milésima. Se tiene:

ZONA RURAL			ZONA URBANA	
Unidad muestral	M_i	Total	Unidad muestral	Total
1	7	13	1	21
2	6	11	2	15
3	8	18	3	24
4	4	10	4	20
5	5	11		

1) Estimar la media por conglomerado en cada zona y sus errores absoluto y relativo de muestreo. Hallar también un intervalo de confianza del 95% para la media por conglomerado en cada zona.

2) Estimar el total en la población y sus errores absoluto y relativo de muestreo.

Comenzaremos por la zona rural, en la cual tenemos definido muestreo unietápico de conglomerados con probabilidades proporcionales a los tamaños y muestreo con reposición, lo que nos lleva a utilizar el estimador de Hansen y Hurwitz. Tenemos:

$$\hat{\bar{X}}_{HHR} = \frac{1}{M_R}\sum_i^n \frac{X_i}{nP_i} = \frac{1}{M_R}\cdot\frac{1}{n}\sum_i^n \frac{X_i}{M_{iR}/M_R} = \frac{1}{n}\sum_i^n \frac{X_i}{M_{iR}} = \frac{1}{5}\left(\frac{13}{7}+\frac{11}{6}+\frac{18}{8}+\frac{10}{4}+\frac{11}{5}\right) = 2{,}128$$

Para estimar la varianza del estimador de la media utilizamos:

$$\hat{V}(\hat{\bar{X}}_{HHR}) = \frac{1}{M_R^2}\hat{V}(\hat{X}_{HHR}) = \frac{1}{M_R^2}\frac{\sum_{i=1}^n\left(\frac{X_i}{P_i}-\hat{X}_{HHR}\right)^2}{n(n-1)} = \frac{1}{M_R^2}\frac{\sum_{i=1}^n\left(\frac{X_i}{M_{iR}/M_R}-M_R\hat{\bar{X}}_{HHR}\right)^2}{n(n-1)} =$$

$$\frac{\sum_{i=1}^n\left(\frac{X_i}{M_{iR}}-\hat{\bar{X}}_{HHR}\right)^2}{n(n-1)} = \frac{\left(\frac{13}{7}-2{,}128\right)^2+\left(\frac{11}{6}-2{,}128\right)^2+\left(\frac{18}{8}-2{,}128\right)^2+\left(\frac{10}{4}-2{,}128\right)^2+\left(\frac{11}{5}-2{,}128\right)^2}{20} = 0{,}016$$

El error relativo de muestreo en la zona rural será:

$$\hat{C}v(\hat{\bar{X}}_{HHR}) = \frac{\sqrt{V(\hat{\bar{X}}_{HHR})}}{\hat{\bar{X}}_{HHR}} = \frac{\sqrt{0{,}016}}{2{,}128} = 0{,}059 \cong 6\%$$

Un intervalo de confianza al 95% para el gasto medio por hogar en zona rural es:

$$\hat{\bar{X}}_{HHR} \pm \lambda_\alpha\sqrt{V(\hat{\bar{X}}_{HHR})} = 2{,}128 \pm 1{,}96\sqrt{0{,}016} = [1{,}880,\quad 2{,}376]$$

Nos ocupamos ahora de la zona urbana, en la cual tenemos definido muestreo sistemático con un coeficiente de correlación intramuestral muy pequeño, lo que nos va a permitir estimar la varianza mediante la fórmula del muestreo aleatorio simple. Tenemos entonces los siguientes estimadores:

$$\hat{\bar{X}}_U = \frac{21+15+24+20}{4} = 20$$

$$V(\hat{\bar{X}}_U) = (1-f)\frac{\hat{S}^2}{n} = \left(1-\frac{4}{40}\right)\frac{\frac{1}{3}\left[(21-20)^2+(15-20)^2+(24-20)^2+(20-20)^2\right]}{4} = 3{,}15$$

El error relativo de muestreo en la zona urbana será:

$$\hat{C}v(\hat{\bar{X}}_U) = \frac{\sqrt{V(\hat{\bar{X}}_U)}}{\hat{\bar{X}}_U} = \frac{\sqrt{3{,}15}}{20} = 0{,}0887 \cong 8{,}87\%$$

Un intervalo de confianza al 95% para el gasto medio por hogar en zona urbana es:

$$\hat{\bar{X}}_U \pm \lambda_\alpha \sqrt{V(\hat{\bar{X}}_U)} = 20 \pm 1{,}96\sqrt{3{,}15} = [16{,}5214,\quad 23{,}4786]$$

Para estimar el total de la población utilizamos el muestreo estratificado, que es el definido en primera etapa, teniendo en cuenta que en segunda etapa están definidos muestreo unietápico de conglomerados en la zona rural, y muestreo sistemático en la zona urbana. Tenemos:

$$\hat{X}_{st} = \sum_{h=1}^{n} N_h \bar{x}_h = 60\hat{\bar{X}}_{HHR} + 40\hat{\bar{X}}_U = 60 \cdot 2{,}128 + 40 \cdot 20 = 927{,}68$$

$$V(\hat{X}_{st}) = \sum_{h=1}^{n} N_h^2 V(\bar{x}_h) = 60^2 V(\hat{\bar{X}}_{HHR}) + 40^2 V(\hat{\bar{X}}_U) = 60^2 \cdot 0{,}016 + 40^2 \cdot 3{,}15 = 5097{,}6$$

$$\hat{C}v(\hat{X}_{st}) = \frac{\sqrt{V(\hat{X}_{st})}}{\hat{X}_{st}} = \frac{\sqrt{5097{,}6}}{927{,}68} = 0{,}077 \cong 7{,}7\%$$

8.7. En las 10 regiones de un país se efectúa muestreo en dos etapas (1ª etapa con reposición). En la primera etapa se obtienen tres regiones de 50, 60 y 80 distritos. En la segunda etapa se seleccionan cinco distritos de cada región de la primera etapa en los que se mide el número de habitantes condenados a cadena perpetua, y se obtienen los siguientes datos:

Unidades primarias de la muestra ($n = 3$)	*Tamaños* (M_i)	*Valores observados* X_{ij} $m_i = \bar{m} = 5$
$REGIÓN_1$	50	8, 6, 12, 14, 10
$REGIÓN_2$	60	8, 10, 14, 14, 16
$REGIÓN_3$	80	8, 10, 10, 16, 12

Sabiendo que el total de distritos es M = 600, se pide formar un estimador insesgado del total X de condenados a cadena perpetua y calcular el valor particular correspondiente a los datos del problema en los siguientes casos:

1) Muestreo con probabilidades iguales en las dos etapas.
2) Muestreo con probabilidades proporcionales al tamaño en primera etapa.
3) Estimar el error de muestreo en ambos casos.

Para probabilidades iguales en ambas etapas el estimador del total es:

$$\hat{X} = \frac{N}{n}\sum_{i}^{n} M_i \bar{x}_i = \frac{10}{3}\left(50 \cdot 10 + 60 \cdot 12{,}4 + 80 \cdot 11{,}2\right) = 7133{,}33 \approx 7134 \text{ condenados}$$

La estimación de la varianza es:

$$\hat{V}\left(\hat{X}\right) = \frac{\sum_{i}^{n}\left(\frac{\hat{X}_i}{1/N} - \hat{X}\right)^2}{n(n-1)} = \frac{\sum_{i}^{n}\left(N\hat{X}_i - N\frac{1}{n}\sum_{i}^{n} M_i\bar{x}_i\right)^2}{n(n-1)} = \frac{N^2}{n}\frac{\sum_{i}^{n}\left(M_i\bar{x}_i - \frac{1}{n}\sum_{i}^{n} M_i\bar{x}_i\right)^2}{n-1} = \frac{N^2}{n}\frac{\sum_{i}^{n}\left(\hat{X}_i - \hat{\bar{X}}_i\right)^2}{n-1}$$

$$\frac{100}{3}\left(\frac{(50\cdot 10 - 713{,}33)^2 + (60\cdot 12{,}4 - 713{,}33)^2 + (80\cdot 11{,}2 - 713{,}33)^2}{2}\right) = 2.19385\cdot 10^7$$

Para probabilidades proporcionales a los tamaños en primera etapa se tiene:

$$\hat{\hat{X}}_{HH} = \frac{1}{n}\sum_i^n \frac{M_i \bar{x}_i}{P_i} = \frac{1}{n}\sum_i^n \frac{M_i \bar{x}_i}{M_i/M} = \frac{M}{n}\sum_i^n \bar{x}_i = \frac{600}{3}(10+12{,}4+11{,}2) = 6720 \text{ condenados}$$

La estimación de la varianza es:

$$\hat{V}(\hat{X}) = \frac{\sum_i^n \left(\frac{\hat{X}_i}{M_i/M} - \hat{X}\right)^2}{n(n-1)} = \frac{\sum_i^n \left(\frac{M}{M_i} M_i \bar{x}_i - \frac{M}{n}\sum_i^n \bar{x}_i\right)^2}{n(n-1)} = \frac{M^2 \sum_i^n \left(\bar{x}_i - \frac{1}{n}\sum_i^n \bar{x}_i\right)^2}{n(n-1)} =$$

$$\frac{600^2\left((10-11{,}2)^2 + (12{,}4-11{,}2)^2 + (11{,}2-11{,}2)^2\right)}{6} = 172800$$

Se observa que el error de muestreo es mucho menor en el caso de utilizar probabilidades proporcionales a los tamaños.

8.8. Consideramos las 1100 granjas de cerdos de una comarca que se estratifican formando 2 estratos. El primero de ellos (granjas en zona rural) tiene 1.000 granjas de 50 cerdos con 4 meses de edad del que se extrae una muestra de 5 granjas, en cada una de las cuales se obtiene a su vez una submuestra de 6 cerdos. Los pesos promedios (en arrobas) de los 6 cerdos con 4 meses de las 5 granjas anteriores extraídas del primer estrato son los siguientes: $\bar{x}_{i1} = \{3, 5, 2, 4, 6\}$ $i = 1, 2, ..., 5$ y $S_{1w}^2 = 1{,}5$. El segundo estrato (granjas en perímetro urbano) tiene 100 granjas de 40 cerdos con 4 meses cada una del que se extrae una muestra de 6 granjas, en cada una de las cuales se obtiene a su vez una submuestra de 4 cerdos. Los pesos promedios (en arrobas) de los 4 cerdos con 4 meses de las 6 granjas anteriores extraídas del segundo estrato son los siguientes: $\bar{x}_{i2} = \{3, 4, 3, 5, 3, 3\}$ $i = 1, 2, ..., 6$ y $S_{2w}^2 = 1{,}33$. A partir de esta información, estimar el peso promedio de los cerdos a los 4 meses en las granjas de la comarca y sus errores absoluto y relativo de muestreo considerando muestreo sin reposición y probabilidades iguales en todas las etapas. Hallar también un intervalo de confianza para el peso promedio de los cerdos a los 4 meses en las granjas de la comarca al 95%.

Estamos ante el típico diseño complejo de muestreo bietápico de conglomerados (granjas de cerdos) con estratificación de las unidades de primera etapa (las granjas) en dos estratos. Las unidades elementales de segunda etapa son los cerdos con 4 meses de las granjas.

Inicialmente estimamos la media y su varianza en el primer estrato. Tenemos:

$$\bar{\bar{x}}_1 = \frac{1}{n_1}\sum_i \bar{x}_{i1} = \frac{20}{5} = 4 \qquad \hat{S}_b^2 = \frac{\bar{m}_1 \sum_i^5 \left(\bar{x}_{i1} - \bar{\bar{x}}_1\right)^2}{n_1 - 1} = 15$$

$$\hat{V}(\bar{\bar{x}}_1) = (1-f_{11})\frac{\hat{S}_{1b}^2}{n_1 \bar{m}_1} + f_{11}(1-f_{12}) \cdot \frac{\hat{S}_{1w}^2}{n_1 \bar{m}_1} = \left(1 - \frac{5}{1000}\right)\frac{15}{30} + \frac{5}{1000}\left(1 - \frac{6}{50}\right) \cdot \frac{1{,}5}{30} = 0{,}5$$

Ahora estimamos la media y su varianza en el segundo estrato. Tenemos:

$$\bar{\bar{x}}_2 = \frac{1}{n_2}\sum_i \bar{x}_{i2} = \frac{21}{6} = 3{,}5 \qquad \hat{S}^2_{2b} = \frac{\bar{m}_2 \sum_i^6 \left(\bar{x}_{i2} - \bar{\bar{x}}_2\right)^2}{n_2 - 1} = 2{,}8$$

$$\hat{V}\left(\bar{\bar{x}}_2\right) = \left(1 - f_{21}\right)\frac{\hat{S}^2_{2b}}{n_2 \bar{m}_2} + f_{21}\left(1 - f_{22}\right)\cdot\frac{\hat{S}^2_{2w}}{n_2 \bar{m}_2} = \left(1 - \frac{6}{100}\right)\frac{2{,}8}{24} + \frac{6}{100}\left(1 - \frac{4}{40}\right)\cdot\frac{1{,}33}{24} = 0{,}113$$

El estimador de la media estratificado será:

$$\bar{x}_{st} = \sum_{h=1}^{2} W_h \bar{\bar{x}}_h = W_1 \bar{\bar{x}}_1 + W_2 \bar{\bar{x}}_2 = \frac{1000}{1100}\cdot 4 + \frac{100}{1100}\cdot 3{,}5 = 3{,}685 \text{ arrobas}$$

La estimación de la varianza del estimador de la media valdrá:

$$\hat{V}(\bar{x}_{st}) = \sum_{h=1}^{2} W_h^2 \hat{V}(\bar{\bar{x}}_h) = W_1^2 \hat{V}(\bar{\bar{x}})_1 + W_2^2 \hat{V}(\bar{\bar{x}}_2) = \left(\frac{1000}{1100}\right)^2 \cdot 0{,}5 + \left(\frac{100}{1100}\right)^2 \cdot 0{,}113 = 0{,}415$$

El error relativo de muestreo se estimará mediante:

$$\hat{C}v(\bar{x}_{st}) = \frac{\sqrt{\hat{V}(\bar{x}_{st})}}{\bar{x}_{st}} = \frac{\sqrt{0{,}415}}{3{,}685} = 0{,}1748 \quad (17{,}48\%)$$

El intervalo de confianza al 95%, suponiendo normalidad, será:

$$\bar{x}_{st} \pm \lambda_\alpha \sqrt{\hat{V}(\bar{x}_{st})} = 3{,}685 \pm 1{,}96\sqrt{0{,}415} = \left[2{,}42,\ 4{,}95\right]$$

8.9. Una empresa tiene que realizar una encuesta en la que las unidades primarias de muestreo son las secciones censales y las unidades de segunda etapa son las familias pertenecientes a las secciones censales. La empresa dispone de agentes entrevistadores que residen en la capital de cada provincia en la que tiene sucursales. Se supone que el coste de enviar un agente a una sección censal es de 500 euros y el de realizar una entrevista a una familia es de 50 euros.

Si existe un presupuesto de 3000000 de euros para realizar la encuesta siendo la característica a estimar la proporción de población activa respecto del total, y por encuestas anteriores se tiene una estimación de dicha proporción del 38% y una estimación del coeficiente de correlación intraconglomerados de 0,05, se pide:

1) Considerando muestreo con reposición, plantear el problema de Lagrange que permite calcular el número óptimo de secciones censales y el de familias a entrevistar dentro de cada una.

2) Hallar el valor de los números óptimos citados para el coste total dado.

Para plantear el problema de Lagrange adecuado, consideramos la función de coste de campo $C = c_1 n + c_2 n\overline{m}$ donde $c_1 = 500$ es el coste de enviar un agente a una sección censal y $c_2 = 50$ es el coste de realizar una entrevista a una familia en segunda etapa. Como el presupuesto total para realizar la encuesta es de 3000000 de euros, la función de coste será:

$$3000000 = 500n + 50n\overline{m}$$

Como la característica a estimar es el porcentaje de población activa respecto del total, utilizaremos la varianza de la proporción para denotar el error, es decir:

$$V(\hat{P}) = (1 - f)\frac{\hat{P}\hat{Q}}{n\overline{m}}(1 + (\overline{m} - 1)\delta)$$

El problema se resuelve minimizando la varianza para el coste dada a través del problema de optimización de Lagrange:

$$\left.\begin{array}{l} Min V(\hat{P}) = (1-f)\dfrac{0{,}38(1.0{,}38)}{n\overline{m}}(1+(\overline{m}-1)0{,}05) \\ 3000000 = 500n + 50n\overline{m} \end{array}\right\} \Rightarrow \overline{m} = \sqrt{\frac{c_1}{c_2}\cdot\frac{1-\delta}{\delta}} = \sqrt{\frac{500}{50}\cdot\frac{1-0{,}05}{0{,}05}} \cong 14 \text{ familias}$$

$$3000000 = 500n + 50n\overline{m} \Rightarrow n = \frac{3000000}{500 + 50\overline{m}} = \frac{3000000}{500 + 50\cdot 14} = 2500 \text{ secciones censales}$$

8.10. Una empresa quiere estimar la proporción de máquinas que han sido retiradas del proceso de producción debido a reparaciones mayores. Para ello utiliza muestreo en dos etapas considerando unidades de primera etapa las plantas de que dispone y unidades de segunda etapa las máquinas de las plantas. Se dispone de tiempo y dinero para muestrear 10 plantas y se obtiene que los tamaños de las plantas M_i, las máquinas muestreadas en cada planta en segunda etapa m_i y las proporciones muestrales de máquinas que requieren reparaciones mayores son los que se exponen en la siguiente tabla:

Planta	M_i	m_i	*Porcentaje de máquinas con reparaciones mayores* $(\hat{P}_i)$
1	50	10	0,40
2	65	13	0,38
3	45	9	0,22
4	48	10	0,30
5	52	10	0,50
6	58	12	0,25
7	42	8	0,38
8	66	13	0,31
9	40	8	0,25
10	56	11	0,36

Estimar la proporción de máquinas que han sido retiradas del proceso de producción debido a reparaciones mayores para todas las plantas y establecer un límite para el error de estimación al 95%.

Al no conocerse el valor M se utilizará el estimador de la proporción por razón al tamaño siguiente:

$$\hat{P} = \frac{\sum_{i=1}^{n} M_i \hat{P}_i}{\sum_{i=1}^{n} M_i} = 0{,}34$$

cuyo error de muestreo puede estimarse mediante:

$$\hat{V}\left(\hat{P}\right) = \frac{\left(1 - f_1\right)}{n\overline{M}^2} \cdot \frac{\sum_{i}^{n} M_i^2 \left(\hat{P}_i - \hat{P}\right)^2}{n-1} + \frac{1}{nN\overline{M}^2} \sum_{i}^{n} M_i^2 \left(1 - f_{2i}\right) \cdot \frac{\hat{P}_i \hat{Q}_i}{m_i - 1} = 0{,}0081$$

Un límite para el error de estimación al 95% será:

$$\hat{P} \pm 2\sqrt{\hat{V}\left(\hat{P}\right)} = 0{,}34 \pm 0{,}056$$

Se estima entonces que la proporción de máquinas involucradas en reparaciones mayores es de 0,34, con un límite para el error de estimación de 0,056.

EJERCICIOS PROPUESTOS

8.1. Se desea estimar el consumo de los hogares españoles a través de una muestra bietápica formada por conglomerados de 500 hogares cuya unidad primaria de muestreo es la sección censal. El coeficiente de correlación intraconglomerados es 0,1. El coste de preparación de listados y planimetría de cada sección censal a incluir en la muestra es de 5.000 unidades monetarias, y el coste de entrevista por hogar es de 1000 unidades monetarias, no considerándose más componentes en la función de coste total. Si se dispone de un presupuesto global de 10000000 de unidades monetarias, se pide:

1) Especificar la función de coste total y plantear el problema de optimización con restricciones asociado.

2) ¿Cuáles serían los tamaños de muestra en cada etapa que optimizasen el diseño? Se entiende por diseño óptimo aquel que logra la máxima precisión dentro del presupuesto fijado.

3) Si se estratifican las secciones censales en dos estratos del mismo tamaño correspondientes a zona rural y zona urbana, de modo que la variabilidad del consumo de los hogares medida a través de la varianza es tres veces superior en la zona urbana que en la rural, ¿cómo se distribuiría la muestra en cada estrato y en cada etapa para optimizar el diseño?

8.2. Un investigador desea muestrear tres hospitales de entre los seis que existen en una ciudad, con el propósito de estimar la proporción de pacientes que han estado (o estarán) en el hospital por más de dos días consecutivos. Puesto que los hospitales varían en tamaño, éstos serán muestreados con probabilidades proporcionales al número de sus pacientes. En los tres hospitales muestreados se examinará un 10% de los registros de los pacientes actuales para determinar cuántos pacientes permanecerán por más de dos días en el hospital. Con la información sobre los tamaños de los hospitales dada en la tabla adjunta se selecciona una muestra de tres hospitales con probabilidades proporcionales al tamaño.

Hosp.	Pacien.	Interv.	Hosp.	Pacien.	Interv.	Hosp.	Pacien.	Interv
1	328	1-328	2	109	329-437	3	432	438-869
4	220	870-1089	5	280	1090-1369	6	190	1370-1559

Puesto que serán seleccionados tres hospitales, tres números aleatorios entre el 0001 y el 1559 deben ser seleccionados de la tabla de números aleatorios. Nuestros números elegidos son 1505, 1256 y 0827. ¿Qué hospitales serán elegidos para la muestra? Supóngase que los hospitales muestreados dieron los siguientes datos sobre el número de pacientes con permanencia de más de dos días:

Hospital	Nº de pacientes muestreados	Nº con más de dos días de permanencia
a	43	25
b	28	15
c	19	8

Estimar la proporción de pacientes con permanencia superior a dos días para los seis hospitales y establecer un límite para el error de estimación.

8.3. Supongamos que cinco investigadores toman muestras independientes de igual tamaño constituidas por pequeñas parcelas de un campo de cultivo y obtienen estimaciones del rendimiento del campo θ. Sean estas estimaciones: 97, 96, 100, 98, 94. Si tomamos como estimador de θ la media de las cinco estimaciones, calcular el error de muestreo relativo. Realizar el mismo cálculo suponiendo que las muestras son de distintos tamaños, de 3, 1, 10, 10 y 1, respectivamente

8.4. Realizamos muestreo bietápico en una población de 10 conglomerados de tamaños desiguales. En la primera etapa se toman tres unidades primarias y en la segunda etapa se toman cinco unidades dentro de cada unidad primaria. Hallar el estimador lineal insesgado del total poblacional en el caso de muestreo sin reposición con probabilidades iguales en las dos etapas. Probar que si se aplica el teorema de Durbin para la estimación de la varianza del estimador del total se tiene:

$$\hat{V}(\hat{X}) = \frac{14}{45}\sum_{i=1}^{3} M_i^2 x_i^2 - \frac{2}{3}\sum_{i=1}^{3} s_i^2 M_i\,(M_i - 5) - \frac{7}{45}\sum_{i\neq j} M_i M_j x_i x_j$$

siendo x_i el total muestral y $s_i^2 = \hat{S}_i^2$ la cuasivarianza dentro de la unidad primaria i-ésima de la muestra. Si consideramos muestreo con reposición en la segunda etapa, ¿cuál es el estimador del total? ¿Qué expresión toma el estimador de su varianza?

8.5. Una cadena de supermercados tiene tiendas en 32 ciudades. Un director de la compañía quiere estimar la proporción de tiendas en la cadena que no satisfacen un criterio de limpieza específico. Las tiendas dentro de cada ciudad poseen características similares, por lo que el director selecciona una muestra por conglomerados en dos etapas que conmtiene la mitad de las tiendas dentro de cada una de las cuatro ciudades. La tabla siguiente muestra los datos recogidos.

Ciudad	*Nº de tiendas en la ciudad*	*Nº de tiendas muestreadas*	*Nº de tiendas que no satisfacen el criterio de limpieza*
1	25	13	3
2	10	5	1
3	18	9	4
4	16	8	2

Estimar la proporción de tiendas que no satisfacen el criterio de limpieza y establecer un límite para el error de estimación al 95% de confianza.

CAPÍTULO

9

MUESTREO BIFÁSICO Y MUESTREO EN OCASIONES SUCESIVAS

OBJETIVOS

1. Presentar el concepto de muestreo bifásico.
2. Analizar los estimadores y sus errores en muestreo bifásico con estratificación.
3. Analizar los estimadores y sus errores en muestreo bifásico para estimaciones de razón.
4. Analizar los estimadores y sus errores en muestreo bifásico para estimaciones de regresión.
5. Analizar los estimadores y sus errores en muestreo bifásico para estimaciones de diferencia.
6. Estudiar los estimadores de mínima varianza en el muestreo en ocasiones sucesivas.

ÍNDICE

1. Muestreo bifásico.

2. Muestreo bifásico para estratificación. Estimadores, varianzas y estimación de varianzas.

3. Muestreo bifásico para estimadores de razón.

4. Muestreo bifásico para estimadores de regresión.

5. Muestreo bifásico para estimadores de diferencia.

6. Mestreo en ocasiones sucesivas.

7. Estimadores de mínima varianza en el muestreo en ocasiones sucesivas.

8. Problemas resueltos.

9. Ejercicios propuestos.

MUESTREO BIFÁSICO

El muestreo doble o bifásico se utiliza cuando queremos obtener estimadores de alguna variable X y disponemos de información adicional de otra variable de modo similar a lo que ocurría en los métodos de estimación indirecta. En la práctica, el muestreo doble se lleva a cabo seleccionando en una primera fase una muestra, relativamente grande, en la que a bajo coste pueden observarse una o varias características generales de las unidades que nos proporcionan la información que necesitamos para el estudio de nuestra característica objetivo. En una segunda fase seleccionamos una submuestra de la primera en la que observamos ya la característica objeto de estimación. Esta técnica se conoce con el nombre de muestreo en dos fases, muestreo doble o muestreo bifásico. Para fijar notación consideramos:

1ª fase. Se toma una muestra grande de tamaño n' relativa a la variable auxiliar Y_i para estimar por ejemplo $\overline{Y}$ u otras características relativas a la variable Y_i con bajo coste.

2ª fase. Se toma una muestra relativa a la variable en estudio X_i de tamaño n (generalmente submuestra de la muestra preliminar $n< n'$) con coste mucho más alto.

El uso de esta técnica de muestreo depende de los costes. Si la observación de la característica X_i que nos interesa no tiene coste, o es muy bajo, sencillamente tomaríamos una muestra del tamaño n_o necesario para la precisión deseada y con ella haríamos las estimaciones relativas a X_i. Supongamos que disponemos de un presuspuesto total C, que el coste por unidad de la primera muestra, de tamaño n', es c' y que el coste por unidad de la segunda muestra, de tamaño $n < n'$, es c. Frecuentemente c' es mucho más pequeño que c, bien sea porque la primera muestra se utiliza para obtener unos pocos datos generales de las unidades (en campo o en oficina, si se dispone de un fichero o registro) o bien porque la observación de la característica objetivo implica un proceso de observación más costoso. En estas condiciones, si tomamos una sola muestra, tendremos $C = cn_o$, y si hacemos muestreo en dos fases $C = c'n' + cn$. Supongamos que los costes totales por el procedimiento bifásico y por el normal (aleatorio) son los mismos, esto es, $cn_o = c'n + cn$. Igualando los dos costes totales, se obtiene: $n_o = n + \frac{c'}{c} n'$, lo que nos dice que con la técnica de dos fases la observación efectiva (la referida a la variable X_i) se hace en una muestra de tamaño n, menor que el tamaño n_o de la muestra aleatoria simple correspondiente en una sola fase con el mismo coste total. Luego al introducir las dos fases el tamaño de muestra necesario es más pequeño que si hubiese una sola fase (muestreo aleatorio normal) y hay una pérdida en la precisión de los estimadores (al disminuir el tamaño de la muestra).

Se trata de decidir si compensa la disminución del tamaño efectivo de la muestra, con el incremento de información adquirido en la primera fase (lo que provocará pérdida de precisión en las estimaciones relativas a X_i). Para ello debe calcularse la varianza correspondiente a muestreo doble y compararla con la del muestreo en una sola fase $\frac{\sigma^2}{n_o}$ en caso de estimación de la media. Es obvio que cuanto menor sea la relación c'/c más favorable es el muestreo doble. Ello es debido a que $n_o - n = (c' / c)\, n' \Rightarrow$ mientras menor sea c' / c más cerca estará n de n_o y menos disminución habrá del tamaño de muestra comparado el bifásico y el aleatorio simple, siendo la pérdida en precisión de los estimadores menor al introducir el bifásico.

La adecuación del muestreo bifásico depende de si lo que se gana en precisión de los estimadores al introducir la ayuda de la muestra grande compensa la pérdida en precisión debida a la reducción del tamaño de la muestra para estimar X_i, esto es, la ayuda de la variable auxiliar Y_i. La primera muestra de tamaño n' proporciona ciertos datos buenos basados en la variable auxiliar Y_i para que las estimaciones finales (las estimaciones de X_i) sean precisas. Si no hubiese variable auxiliar Y_i el tamaño de la muestra para estimar X_i será n_o, y al introducir la variable auxiliar el tamaño de la muestra sería $n < n_o$.

MUESTREO BIFÁSICO PARA ESTRATIFICACIÓN. ESTIMADORES, VARIANZAS Y ESTIMACIÓN DE VARIANZAS

Partimos de una población estratificacada en L clases (estratos). *La primera muestra (primera fase)* es aleatoria de tamaño n' seleccionada de entre las n unidades de la población. Sea W_h = Proporción de elementos de la población que caen en el estrato h, que es desconocida inicialmente.

$$W_h = \frac{N_h}{N} = \frac{\text{Número de elementos poblacionales en el estrato } h}{\text{Número total de elementos de la población}}$$

Consideremos ahora la proporción de elementos de la primera muestra que cae en el estrato h:

$$\hat{W}_h = \frac{n'_h}{n'} = \frac{\text{Número de elementos de la primera muestra que caen en el estrato } h}{\text{Número total de elementos de la primera muestra}}$$

Hay que tener presente que si consideramos selecciones diferentes de la primera muestra (con n' prefijado) obtenemos diferentes valores de n'_h y $\hat{W}_h$ resulta ser un estimador insesgado de W_h (porque la proporción muestral en muestreo aleatorio simple es un estimador insesgado de la proporción poblacional, lo mismo que la media muestral es un estimador insesgado de la media poblacional). Tenemos entonces que $E\left(\hat{W}_h\right) = W_h$ estando la esperanza referida a las muestras posibles de n' unidades de entre las N de la población. A efectos de clarificar la notación especificamos lo siguiente:

$n'_h = n^{\text{o}}$ de unidades de entre las n' de la muestra de primera fase que caen en el estrato h para $h = 1, 2, ..., L$

$$n' = \sum_{h=1}^{L} n'_h \quad \text{y} \quad n = \sum_{h=1}^{L} n_h$$

La segunda muestra (segunda fase) es una muestra aleatoria estratificada de tamaño n. Consiste en tomar una submuestra aleatoria de tamaño $n_h \le n'_h$ en cada estrato independientemente (o sea, las n_h las elegimos de entre las n'_h para valores de $h = 1, ..., L$). Tendremos $n = \sum_{h=1}^{L} n_h$. Ahora n' es dado y n'_1 ....n'_h ....n'_L son fijos y $\hat{W}_1 \cdots \hat{W}_h \cdots \hat{W}_L$ también serán fijos (por serlo n'_h y n') y lo que se hace es considerar todas las submuestras aleatorias de n_h unidades que pueden extraerse de entre las n'_h unidades dadas.

Estimadores y varianzas

El estimador usual de la media en muestreo estratificado es $\hat{\bar{X}} = \sum_h W_h \bar{x}_h$ con $W_h = \frac{N_h}{N}$. En muestreo doble los W_h se estiman por los $\hat{W}_h$ obtenidos de la primera muestra, y con la segunda muestra estimamos las medias $\bar{x}_h = \frac{x_h}{n_h}$; de esta forma resulta el estimador para la media:

$$\hat{\bar{X}} = \sum_h \hat{W}_h \bar{x}_h \quad ; \quad \hat{W}_h = \frac{n'_h}{n'}$$

Utilizaremos la notación $E_{W'}(T)$ para expresar la esperanza matemática de un estadístico *T*, condicionada al conjunto de muestras de primera fase en las cuales *n'*$_1$, ...,*n'*$_h$, ..., *n'* son fijos, o lo que es lo mismo, para un *n'* dado, $\hat{W}_1, \cdots, \hat{W}_h, \cdots, \hat{W}_L$ son fijos. Análogamente $V_{W'}(T)$ expresará la varianza condicionada.

La varianza del estimador de la media *sin reposición en las dos fases* es:

$$V\left(\hat{\bar{X}}\right) = \sum_h (1 - f_h) \frac{S_h^2}{n_h} \left(W_h^2 + \frac{g' W_h (1 - W_h)}{n'} \right) + \frac{g'}{n'} \sum_h W_h \left(\bar{X}_h - \bar{X} \right)^2$$

donde *g'* es el factor de finitud *g'* = (*N* – *n'*)/(*N* –1). Por otro lado, Rao expresó esta varianza de la media de la siguiente forma:

$$V\left(\hat{\bar{X}}\right) = \frac{N - n'}{N} \cdot \frac{S^2}{n'} + \sum_h \left(\frac{1}{\nu_h} - 1 \right) \cdot W_h \frac{S_h^2}{n'} \quad ; \quad \nu_h = \frac{n_h}{n'_h}$$

Para muestreo es *con reposición en primera fase* tendremos:

$$V\left(\hat{\bar{X}}\right) = \sum_h (1 - f_h) \frac{S_h^2}{n_h} \left(W_h^2 + \frac{W_h (1 - W_h)}{n'} \right) + \frac{1}{n'} \sum_h W_h \left(\bar{X}_h - \bar{X} \right)^2$$

fórmula aproximada para *n'* pequeño respecto de *N* en *caso sin reposición en segunda fase*.

Para muestreo *con reposición en las dos fases* tendremos:

$$V\left(\hat{\bar{X}}\right) = \sum_h \frac{\sigma_h^2}{n_h} \left(W_h^2 + \frac{W_h (1 - W_h)}{n'} \right) + \frac{1}{n'} \sum_h W_h \left(\bar{X}_h - \bar{X} \right)^2$$

fórmula aproximada para n_h pequeño respecto de N_h, en todo *h*, y *n'* pequeño respecto de *N*.

Para el total $X = N\bar{X}$, el estimador insesgado es $\hat{X} = N\hat{\bar{X}}$ y su varianza es $V\left(\hat{X}\right) = N^2 V\left(\hat{\bar{X}}\right)$.

Si la muestra de primera fase es de tamaño $n'=N$, esto es, se observan todas las unidades de la población para efectuar la estratificación, la fórmula general de la varianza del estimador en muestreo doble se convierte en:

$$V\left(\hat{\bar{X}}\right) = \sum_h \left(1 - f_h\right) W_h^2 \frac{S_h^2}{n_h} \quad ; \quad g' = 0$$

que coincide con la del muestreo estratificado habitual (una sola fase). Además se observa que n' aparece dividiendo, y en consecuencia, cuanto mayor es n' ($n' < N$) la pérdida de precisión por el uso de muestreo doble disminuye. Obviamente el coste aumenta, razón por la cual conviene estudiar los tamaños y la afijación óptimos en función del coste.

Para *proporciones y totales de clase* tenemos:

Si se desea estimar una porporción P en la población, siendo P_h la correspondiente al h-ésimo estrato, el estimador insesgado en muestreo doble es:

$$\hat{P} = \sum_h \hat{W}_h p_h \quad ; \quad p_h = \text{proporción muestral en segunda fase.}$$

La varianza (*sin reposición en las dos fases*), aplicando el resultado anterior, será:

$$V\left(\hat{P}\right) = \sum_h \left(1 - f_h\right) \frac{P_h Q_h}{n_h} \left(W_h^2 + \frac{g' W_h \left(1 - W_h\right)}{n'} \right) + \frac{g'}{n'} \sum_h W_h \left(P_h - P\right)^2$$

con la aproximación $S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h \approx P_h Q_h$.

En *muestreo con reposición en las dos fases*, o sin reposición y tamaños muestrales pequeños respecto de los correspondientes poblacionales $\left(f_h \approx 1;\ g' \approx 1\right)$, se tiene:

$$V\left(\hat{P}\right) = \sum_h \frac{P_h Q_h}{n_h} \left(W_h^2 + \frac{W_h \left(1 - W_h\right)}{n'} \right) + \frac{1}{n'} \sum_h W_h \left(P_h - P\right)^2$$

Para el total de clase, $A = NP$, el estimador es $\hat{A} = N\hat{P}$ y su varianza $V\left(\hat{A}\right) = N^2 V\left(\hat{P}\right)$.

Para *afijación proporcional*, si en la muestra de segunda fase asignamos a cada estrato un tamaño muestral n_h proporcional al tamaño del estrato, se tiene $n_h = W_h n$, resultando para la varianza del estimador la fórmula:

$$V\left(\hat{\bar{X}}\right) = \frac{1}{n} \sum_h \left(1 - f_h\right) S_h^2 \left(W_h + \frac{g' \left(1 - W_h\right)}{n'} \right) + \frac{g'}{n'} \sum_h W_h \left(\bar{X}_h - \bar{X}\right)^2$$

En la práctica, para efectuar la afijación a los estratos utilizaremos $n_h = \hat{W}_h n$.

En muestreo con reposición se tiene:

$$V\left(\hat{\bar{X}}\right) = \frac{1}{n}\sum_h \sigma_h^2 W_h + \frac{1}{nn'}\sum_h \sigma_h^2\left(1 - W_h\right) + \frac{1}{n'}\sum_h W_h\left(\bar{X}_h - \bar{X}\right)^2$$

que puede aproximarse por:

$$V\left(\hat{\bar{X}}\right) = \frac{1}{n}\sum_h W_h \sigma_h^2 + \frac{1}{n'}\sum_h W_h\left(\bar{X}_h - \bar{X}\right)^2$$

Para afijación óptima tenemos:

$$V\left(\hat{\bar{X}}\right) = \frac{1}{n}\left(\sum_h W_h \sigma_h\right)^2 + \frac{1}{n'}\sum_h W_h\left(\bar{X}_h - \bar{X}\right)^2$$

Además, para determinar los tamaños óptimos *n'* y *n* correspondientes a un coste total dado tales que $V\left(\hat{\bar{X}}\right)$ sea mínima, escribimos la función de Lagrange:

$$\phi = \frac{1}{n}A + \frac{1}{n'}B + \lambda(c'n' + cn - C) \text{ con } A = \left(\sum_h W_h \sigma_h\right)^2 \text{ y } B = \sum_h W_h\left(\bar{X}_h - \bar{X}\right)^2$$

Derivando respecto de *n* y *n'* y λ se tiene:

$$\left.\begin{aligned} \frac{\partial\phi}{\partial n} &= -\frac{A}{n^2} + \lambda c = 0 \Rightarrow \lambda = \frac{A}{cn^2} \\ \frac{\partial\phi}{\partial n'} &= -\frac{B}{n'^2} + \lambda c' = 0 \Rightarrow \lambda = \frac{B}{c'n'^2} \\ \frac{\partial\phi}{\partial\lambda} &= c'n' + cn - C = 0 \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} n &= \frac{C\sqrt{A}}{\sqrt{c}\left(\sqrt{Ac} + \sqrt{Bc'}\right)} \\ n' &= \frac{C\sqrt{B}}{\sqrt{c'}\left(\sqrt{Ac} + \sqrt{Bc'}\right)} \\ V_{\acute{o}pt.}\left(\hat{\bar{X}}\right) &= \frac{\left(\sqrt{Ac} + \sqrt{Bc'}\right)^2}{C} \end{aligned}\right.$$

Estimación de varianzas

Tenemos:

$$\hat{V}\left(\hat{\bar{X}}\right) = \frac{n'}{n'-1}\left[\sum_h \frac{s_h^2}{n_h}\left(\hat{W}_h^2 - \frac{\hat{W}_h}{n'}\right) + \frac{1}{n'}\sum_h \hat{W}_h\left(\bar{x}_n - \bar{X}\right)^2\right]$$

El factor $\dfrac{n'}{(n'-1)}$ prácticamente es próximo a la unidad si *n'* no es pequeño. También el término que aparece en segundo lugar en la fórmula de la estimación de la varianza puede ser despreciable respecto de los otros dos, ya que aparece el producto $n_h \cdot n'$ en el denominador. Entonces resulta la aproximación:

$$\hat{V}\left(\hat{\bar{X}}\right) \approx \sum_h \hat{W}_h^2 \frac{s_h^2}{n_h} + \frac{1}{n'} \sum \hat{W}_h \left(\bar{x}_h - \hat{\bar{X}}\right)^2$$

Y, por último, también en esta expresión el segundo sumando será pequeño respecto del primero para valores grandes de *n'*, resultando como fórmula aproximada más sencilla:

$$\hat{V}\left(\hat{\bar{X}}\right) \approx \sum_h \hat{W}_h^2 \frac{s_h^2}{n_h}$$

que es la correspondiente a muestreo estratificado en una sola fase, sustituyendo W_h por su estimación $\hat{W}_h$.

En caso de estimar la varianza de la proporción $\hat{P}$ o del total de clase $\hat{A}$, sustituimos en la fórmula para la varianza, o en sus aproximaciones, cuando sean válidas, los siguientes valores:

$$\frac{s_h^2}{n_h} = \frac{p_h q_h}{n_h - 1} \quad ; \quad \left(\bar{x}_h - \hat{\bar{X}}\right)^2 = \left(p_h - \hat{P}\right)^2$$

MUESTREO BIFÁSICO PARA ESTIMADORES DE RAZÓN

El estimador usual de razón para la media $\bar{X}$ utiliza como información conocida previamente la media $\bar{Y}$ (o el total) de una característica *Y*, definida en todas las unidades de la población, elegida convenientemente de modo que su relación con *X* sea lineal al menos aproximadamente. El muestreo doble utiliza la primera muestra de tamaño *n'* para obtener una buena estimación de $\bar{Y}$, o de *Y*, y la segunda muestra de tamaño *n* para estimar $\bar{x}$ e $\bar{y}$. De esta forma , el estimador de razón para la media en muestreo doble es:

$$\hat{\bar{X}}_R = \frac{\bar{x}}{\bar{y}} \cdot \bar{y}' \quad ; \quad \bar{y}' = \text{ Media de la primera muestra.}$$

En el caso de que *las muestras de las dos fases sean independientes*, se tiene:

$$V\left(\hat{\bar{X}}_R\right) = \frac{1}{n}\left\{\sigma_x^2 + R^2\sigma_y^2 - 2R\sigma_{xy}\right\} + \frac{1}{n'} R^2 \sigma_y^2$$

fórmula válida para *muestreo con reposición*. En el caso *sin reposición* sustituimos varianzas y covarianzas por cuasivarianzas y cuasicovarianzas, multiplicando el primer sumando por el factor de finitud en segunda fase y el segundo sumando por el de primera fase.

Para el caso en que *la segunda muestra de tamaño n es una submuestra aleatoria de la primera (n ≤ n')*, resulta:

$$V\left(\hat{\bar{X}}_R\right) = \frac{1}{n}\left\{\sigma_x^2 + R^2\sigma_y^2 - 2R\sigma_{xy}\right\} + \frac{1}{n'}\left\{2R\sigma_{xy} - R^2\sigma_y^2\right\}$$

Para estimar el total en muestreo doble, tendremos:

$$\hat{X}_R = N\hat{\bar{X}}_R \quad ; \quad V\left(\hat{X}_R\right) = N^2 V\left(\hat{\bar{X}}_R\right)$$

Para *estimar la varianza*, dado que en la segunda muestra de tamaño *n* obtenemos observaciones de la variable conjunta (*X,Y*), podemos calcular estimaciones de σ_y^2 y Cov(*X,Y*) como $s_x^2 = \frac{1}{n-1}\sum_1^n (X_i - \bar{x})^2$ y $s_{xy} = \frac{1}{n-1}\sum_1^n (X_i - \bar{x})(Y_i - \bar{y})$, y puesto que la primera muestra es de tamaño *n*' > *n,* nos permite una buena estimación de σ_y^2 mediante $s_y^2 = \frac{1}{n'-1}\sum_1^{n'} (Y_i - \bar{y}')^2$. Para la razón *R*, tomaremos la estimación $\hat{R}$.

MUESTREO BIFÁSICO PARA ESTIMADORES DE REGRESIÓN

El estimador usual para la media en muestreo indirecto (en una fase) por regresión lineal es $\hat{\bar{X}} = \bar{x} + K\left(\bar{Y} - \bar{y}\right)$, donde *K* es una constante prefijada e $\bar{Y}$ es la media poblacional de la variable auxiliar. Los estimadores $\bar{x}, \bar{y}$ se obtienen de las observaciones de una muestra $\left(X_i, Y_i\right)$ de tamaño *n*. En muestreo doble, al suponer desconocida $\bar{Y}$, utilizamos la primera muestra de tamaño *n'* para estimar $\bar{Y}$, estimación dada por $\bar{y}'$. Con la muestra de tamaño *n* en segunda fase estimamos $\bar{x}, \bar{y}$, formando entonces el estimador en muestreo doble por regresión para la media poblacional:

$$\hat{\bar{X}}_{rg} = \bar{x} + K\left(\bar{y}' - \bar{y}\right)$$

En esta situación, la segunda muestra puede ser independiente de la primera o la segunda muestra puede ser una submuestra aleatoria $n < n'$ de la primera.

Si las muestras de las dos fases son independientes, se tiene:

$$V\left(\hat{\bar{X}}_{rg}\right) = \frac{1}{n}\left(\sigma_x^2 + K^2\sigma_y^2 - 2K\sigma_{xy}\right) + \frac{K^2\sigma_y^2}{n'}$$

Para el caso en que *la segunda muestra de tamaño n es una submuestra aleatoria de la primera (n ≤ n')*, resulta:

$$V\left(\hat{\bar{X}}_{rg}\right) = \frac{1}{n}\left(\sigma_x^2 + K^2\sigma_y^2 - 2K\sigma_{xy}\right) + \frac{1}{n'}\left(2K\sigma_{xy} - K^2\sigma_y^2\right)$$

Sea el valor óptimo de $K = b = \frac{\sigma_{xy}}{\sigma_y^2}$ estimado por $\hat{b} = \frac{\sum_1^n (X_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum_1^n (X_i - \bar{x})^2 (Y_i - \bar{y})^2}}$.

Se obtiene en ambos casos (muestras independientes y segunda muestra submuestra de la primera) la expresión para la *varianza óptima del estimador bifásico por regresión*:

$$V\left(\hat{\bar{X}}_{rl}\right)=\frac{\left(1-\rho^2\right)\sigma_x^2}{n}+\frac{\rho^2\sigma_x^2}{n'}-\frac{\sigma_x^2}{N}$$

Una estimación para la varianza óptima es la siguiente:

$$\hat{V}\left(\hat{\bar{X}}_{rl}\right)=\frac{\hat{S}_{x,y}^2}{n}+\frac{\hat{S}_x^2-\hat{S}_{x,y}^2}{n'}-\frac{\hat{S}_x^2}{N}$$

$$\hat{S}_{x,y}^2=\frac{1}{n-2}\left[\sum_{i=1}^{n}(X_i-\bar{x})^2-b^2\sum_{i=1}^{n}(Y_i-\bar{y})^2\right] \qquad \hat{S}_x^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i-\bar{x})^2\right]$$

MUESTREO BIFÁSICO PARA ESTIMADORES DE DIFERENCIA

El estimador por diferencia en muestreo doble resulta del estimador de regresión haciendo $K = 1$, por lo que toda la teoría anterior es válida haciendo K=1, resultando el estimador $\hat{\bar{X}}_d=\bar{x}+\left(\bar{y}'-\bar{y}\right)$. Análogamente, las fórmulas de las varianzas se obtienen aplicando a K el valor 1 en las varianzas del estimador por regresión.

MUESTREO EN OCASIONES SUCESIVAS

El muestreo en ocasiones sucesivas es adecuado cuando estamos interesados en estudiar la evolución de una determinada característica de la población a lo largo del tiempo (como, por ejemplo, la producción industrial, los salarios, la población activa, etc.), para lo que se toman periódicamente muestras del mismo colectivo. En esta situación es habitual que un objetivo sea estimar el cambio producido en la variable estudiada desde la ocasión anterior, otro objetivo puede ser estimar el valor promedio de la media sobre las dos ocasiones, e incluso otro objetivo puede ser estimar la media para la ocasión más reciente.

Inicialmente puede diseñarse una muestra que permanece fija de una ocasión a otra, pero, aunque metodológicamente ésta es la situación más ventajosa, tiene el inconveniente de que las personas o entidades encuestadas son reacias a permanecer por un tiempo indefinido en dicha muestra. Para tratar de resolver este problema se utiliza un procedimiento que consiste en sustituir, en cada período de encuesta, una parte de la muestra, lo que da lugar a la denominada rotación de la muestra. Conviene observar de pasada que esto no siempre puede practicarse, ya que, cuando se trata de unidades muy grandes (grandes almacenes, siderúrgicas, astilleros, etc.), a veces una o unas pocas contribuyen al total estimado en una cantidad superior a todas las demás juntas. En este caso prescindiríamos del muestreo incluyendo estas unidades críticas en un estrato de unidades autorrepresentadas (de probabilidad 1). Adicionalmente surge la pregunta: ¿Con qué frecuencia y de qué manera debería cambiarse la muestra conforme progresa el tiempo?

Otro problema que puede plantearse es el de la estimación óptima de la segunda ocasión, utilizando las informaciones disponibles, tanto de la ocasión presente como de la anterior. En cualquier caso el valor X, que toma la variable en la unidad A, puede cambiar de una ocasión a la siguiente, desempeñando un papel importante en esta teoría el coeficiente de correlación lineal entre los valores de la variable en una y otra ocasión. De todas formas, las unidades de la muestra en una ocasión pueden ser las mismas que en la ocasión anterior, algunas nuevas y otras permanecientes y seleccionadas independientemente de nuevo todas.

Estimación del cambio entre ocasiones sucesivas

Supongamos que se pretende estimar el cambio de la media entre dos ocasiones, que designaremos por t_1 y por t_2, con una muestra de n unidades. Si utilizamos el estimador simple del cambio:

$$\hat{\partial} = \bar{x}_2 - \bar{x}_1 = \frac{1}{n}\sum_i^n \left(x_{2i} - x_{1i} \right)$$

podemos optar entre las siguientes alternativas:

a) Utilizar la misma muestra, denominada ***panel***, en ambas ocasiones.
b) Mantener en la segunda ocasión c unidades de la primera muestra, eliminar $n-c$ y añadir $n-c$ nuevas unidades.
c) Utilizar en la segunda ocasión una muestra independiente de la primera.

La posibilidad a) nos permitiría conocer los cambios individuales entre las dos ocasiones. Este esquema presenta serias dificultades cuando hemos de medir un carácter en ocasiones sucesivas. Prescindiendo del caso en que las mediciones fuesen destructivas, sería muy difícil mantener indefinidamente las mismas unidades, y aun en el caso de que fuese posible no sería deseable por los sesgos que una exposición continuada a los métodos de encuesta pueden originar en la conducta de los entrevistados. En este sentido puede decirse que la muestra se "contamina" con el tiempo.

Para la posibilidad b), si representamos por c el número de unidades comunes, por $n - c = \bar{c}$ el número de las no comunes, y con los subíndices 1 y 2 las correspondientes ocasiones, se puede hacer la representación gráfica siguiente sobre los solapamientos en los totales muestrales en ambas ocasiones.

Ocasión	t_1	$\leftarrow$	$x_{1\bar{c}}$	$\rightarrow$	$\leftarrow$	x_{1c}	$\rightarrow$
Ocasión	t_2				$\leftarrow$	x_{2c}	$\rightarrow$ $\leftarrow$ $x_{2\bar{c}}$ $\rightarrow$

Las medias en ambas ocasiones son:

$$\bar{x}_1 = \frac{x_{1\bar{c}} + x_{1c}}{n} = \frac{x_{1\bar{c}}}{n} + \frac{x_{1c}}{n} = \frac{n-c}{n}\bar{x}_{1\bar{c}} + \frac{c}{n}\bar{x}_{1c}$$

$$\bar{x}_2 = \frac{x_{2\bar{c}} + x_{2c}}{n} = \frac{x_{2\bar{c}}}{n} + \frac{x_{2c}}{n} = \frac{n-c}{n}\bar{x}_{2\bar{c}} + \frac{c}{n}\bar{x}_{2c}$$

y prescindiendo del factor de corrección para poblaciones finitas $1 - f$ y suponiendo por comodidad que la cuasivarianza poblacional en las dos ocasiones es la misma, tendremos para las varianzas y covarianzas las expresiones:

$$V\left(\bar{x}_1\right) = \frac{S^2}{n}, \quad V\left(\bar{x}_2\right) = \frac{S^2}{n}$$

$$\operatorname{cov}\left(\bar{x}_1, \bar{x}_2\right) = \frac{c^2}{n^2} \cdot \operatorname{cov}\left(\bar{x}_{1c}, \bar{x}_{2c}\right) = \rho_{12} \cdot \frac{S}{\sqrt{c}} \cdot \frac{S}{\sqrt{c}} \cdot \frac{c^2}{n^2} = \rho_{12} \cdot \frac{S^2}{n} \cdot \frac{c}{n} = \rho_{12} \cdot \frac{S^2}{n} \cdot \pi_c$$

Sustituyendo estos valores en la varianza de $\hat{\partial}$ tenemos:

$$V\left(\hat{\partial}\right) = V\left(\bar{x}_1\right) + V\left(\bar{x}_2\right) - 2\,\mathrm{cov}\left(\bar{x}_1\bar{x}_2\right) = \frac{S^2}{n} + \frac{S^2}{n} - 2\frac{S^2}{n}\rho_{12}\pi_c = 2\frac{S^2}{n}\left[1 - \rho_{12}\pi_c\right]$$

siendo ρ_{12} el coeficiente de correlación entre los valores comunes a ambas ocasiones y π_c la proporción de unidades comunes. De esta expresión deducimos que para $\rho_{12} > 0$ la ganancia en precisión es proporcional a $\pi_c\rho_{12}$ correspondiendo la máxima ganancia a los valores $\rho_{12} = +1\ y\ \pi_c = 1$. Por lo tanto, la situación ideal es aquella en la que la proporción de unidades comunes en la muestra en las dos ocasiones es del 100% ($\pi_c = 1$), lo que significa que la muestra es común en su totalidad en las dos ocasiones. La situación también es ideal cuando el coeficiente de correlación entre los valores comunes en ambas ocasiones es máximo ($\rho_{12} = +1$), que en términos prácticos significa que las unidades muestrales en las dos ocasiones han de estar muy estrechamente relacionadas de forma positiva (lo mejor es que sean iguales las muestras en las dos ocasiones).

Estimación de la media extendida a dos ocasiones

Uno de los objetivos clásicos en el muestreo en ocasiones sucesivas es estimar el valor promedio de la media sobre las dos ocasiones. Para ello, consideremos el estimador siguiente:

$$\bar{x} = \frac{1}{2}\left(\bar{x}_1 + \bar{x}_2\right)$$

definido como la media de las medias en ambas ocasiones. Su varianza es:

$$V(\bar{x}) = \frac{1}{4}\left[V\left(\bar{x}_1\right) + V\left(\bar{x}_2\right) + 2\,\mathrm{cov}\left(\bar{x}_1, \bar{x}_2\right)\right]$$

y sustituyendo en la fórmula los valores obtenidos en la sección anterior ($V\left(\bar{x}_1\right) = \frac{S^2}{n}$, $V\left(\bar{x}_2\right) = \frac{S^2}{n}$ y $\mathrm{cov}\left(\bar{x}_1, \bar{x}_2\right) = \frac{S^2}{n}\rho_{12}\pi_c$), tenemos:

$$V\left(\bar{x}\right) = \frac{1}{4}\left[\frac{2S^2}{n} + \frac{2S^2}{n}\rho_{12}\pi_c\right] = \frac{S^2}{2n}\cdot\left[1 + \rho_{12}\pi_c\right]$$

Como este valor es mínimo cuando $\pi_c = 0$, vemos que, en el caso $\rho_{12} < 0$, para estimar la media sobre dos ocasiones es preferible utilizar muestras independientes.

ESTIMADORES DE MÍNIMA VARIANZA EN EL MUESTREO EN OCASIONES SUCESIVAS

Estimador del cambio entre dos ocasiones

Consideraremos el *estimador lineal de mínima varianza del cambio* combinado:

$$\hat{\Delta} = W\left(\bar{x}_{2c} - \bar{x}_{1c}\right) + (1-W)\cdot\left(\bar{x}_{2\bar{c}} - \bar{x}_{1\bar{c}}\right)$$

y determinamos el valor de W que haga efectivamente mínima la varianza de $\hat{\Delta}$.

Tenemos $V\left(\hat{\Delta}\right) = W^2 V\left(\bar{x}_{2c} - \bar{x}_{1c}\right) + \left(1-W\right)^2 V\cdot\left(\bar{x}_{2\bar{c}} - \bar{x}_{1\bar{c}}\right)$.

Obteniendo la primera derivada respecto de W e igualando a cero se tiene:

$$2W\cdot V\left(\bar{x}_{2c} - \bar{x}_{1c}\right) - 2\cdot\left(1-W\right)\cdot V\left(\bar{x}_{2\bar{c}} - \bar{x}_{1\bar{c}}\right) = 0 \Rightarrow W = \frac{V\left(\bar{x}_{2\bar{c}} - \bar{x}_{1\bar{c}}\right)}{V\left(\bar{x}_{2c} - \bar{x}_{1c}\right) + V\left(\bar{x}_{2\bar{c}} - \bar{x}_{1\bar{c}}\right)}$$

y sustituyendo las varianzas $V\left(\bar{x}_{2\bar{c}} - \bar{x}_{1\bar{c}}\right) = \frac{2S^2}{n-c}$ y $V\left(\bar{x}_{2c} - \bar{x}_{1c}\right) = \frac{2S^2}{c}\left(1-\rho_{12}\right) \Rightarrow$

$$W = \frac{\frac{1}{n-c}}{\frac{1}{n-c} + \frac{1-\rho_{12}}{c}} = \frac{c}{c + (n-c)(1-\rho_{12})} = \frac{\pi_c}{1-\rho_{12}(1-\pi_c)} \Rightarrow 1-W = \frac{(1-\rho_{12})(1-\pi_c)}{1-\rho_{12}(1-\pi_c)}$$

Sustituyendo estos valores en la expresión de la varianza del estimador lineal de mínima varianza se obtiene:

$$V\left(\hat{\Delta}\right) = W^2 V\left(\bar{x}_{2c} - \bar{x}_{1c}\right) + \left(1-W\right)^2 V\cdot\left(\bar{x}_{2\bar{c}} - \bar{x}_{1\bar{c}}\right) \frac{\pi_c 2S^2(1-\rho_{12})}{\left[1-\rho_{12}(1-\pi_c)\right]^2\cdot n} = \frac{(1-\pi_c)\cdot(1-\rho_{12})^2 2S^2}{\left[1-\rho_{12}(1-\pi_c)\right]^2\cdot n}$$

$$= \frac{2S^2(1-\rho_{12})}{\left[1-\rho_{12}(1-\pi_c)\right]^2\cdot n}\cdot\left[\pi_c + (1-\pi_c)\cdot(1-\rho_{12})\right] = \frac{2S^2(1-\rho_{12})}{\left[1-\rho_{12}(1-\pi_c)\right]^2\cdot n}\cdot\left(1-\rho_{12}+\pi_c\rho_{12}\right)$$

$$= \frac{2S^2(1-\rho_{12})}{\left[1-\rho_{12}(1-\pi_c)\right]^2\cdot n}\cdot\left(1-\rho_{12}(1-\pi_c)\right) = \frac{2S^2(1-\rho_{12})}{\left[1-\rho_{12}(1-\pi_c)\right]\cdot n}$$

Hemos obtenido una *expresión para la varianza mínima del estimador lineal*:

$$\boxed{V\left(\hat{\Delta}\right) = \frac{2S^2(1-\rho_{12})}{\left[1-\rho_{12}(1-\pi_c)\right]\cdot n}}$$

Vemos que, en este caso, ***el estimador lineal de mínima varianza combinado*** $\hat{\Delta}$ ***proporciona igual precisión que el estimador simple*** $\hat{\partial}$ ***cuando*** $\pi_c = 1$***, es decir, cuando se mantiene la misma muestra para la segunda ocasión.***

Estimador de la media en la segunda ocasión

Vamos a trabajar en la suposición de que en la primera ocasión el tamaño de la muestra es lo suficientemente grande para poder considerar la estimación $\bar{x}_1$ como aproximación al valor $\bar{X}_1$ en el estimador de regresión $\bar{x}'_{2c} = \bar{x}_{2c} + b\left(\bar{x}_1 - \bar{x}_{1c}\right)$ cuya varianza viene dada por la varianza de sus componentes $\bar{x}_{2c} - b\bar{x}_{1c}$ y $b\bar{x}_1$:

$$V\left(\bar{x}_{2c} - b\bar{x}_{1c}\right) = V\left(\bar{x}_{2c}\right) + b^2 V\left(\bar{x}_{1c}\right) - 2\operatorname{cov}\left(\bar{x}_{2c}; \bar{x}_{1c}\right) =$$

$$\frac{S^2}{c} + \rho_{12}^2 \frac{S^2}{c} - 2\rho_{12} \cdot \rho_{12} \cdot \frac{S}{\sqrt{c}} \cdot \frac{S}{\sqrt{c}} = \frac{S^2}{c}\left(1 - \rho_{12}^2\right)$$

$$V\left(b\bar{x}_1\right) = b^2 \cdot V\left(\bar{x}_1\right) = b^2 \cdot \frac{S^2}{n} = \rho_{12}^2 \frac{S^2}{n}, \quad \left(S_1 = S_2 \Rightarrow b = \frac{S_1}{S_2} \cdot \rho_{12} = \rho_{12}\right)$$

Sumando ambas componentes se obtiene: $V\left(\bar{x}'_{2c}\right) = S^2\left(\frac{1 - \rho_{12}^2}{c} + \frac{\rho_{12}^2}{n}\right)$

Utilizaremos el *estimador lineal de mínima varianza de la media para la segunda ocasión* combinado definido por:

$$\bar{x}_2 = W\bar{x}'_{2c} + (1 - W)\bar{x}_{2\bar{c}}$$

cuya varianza $V\left(\bar{x}_2\right) = W^2 V\left(\bar{x}'_{2c}\right) + (1 - W)^2 V\left(\bar{x}_{2\bar{c}}\right)$ es mínima para:

$$W = \frac{V\left(\bar{x}_{2\bar{c}}\right)}{V\left(\bar{x}'_{2c}\right) + V\left(\bar{x}_{2\bar{c}}\right)} \qquad 1 - W = \frac{V\left(\bar{x}'_{2c}\right)}{V\left(\bar{x}'_{2c}\right) + V\left(\bar{x}_{2\bar{c}}\right)}$$

de donde se deduce que el estimador combinado de varianza mínima para estimar la media en la segunda ocasión toma la forma:

$$\bar{x}_2 = \frac{\frac{1}{V\left(\bar{x}'_{2c}\right)}}{\frac{1}{V\left(\bar{x}_{2\bar{c}}\right)} + \frac{1}{V\left(\bar{x}'_{2c}\right)}} \cdot \bar{x}'_{2c} + \frac{\frac{1}{V\left(\bar{x}_{2\bar{c}}\right)}}{\frac{1}{V\left(\bar{x}_{2\bar{c}}\right)} + \frac{1}{V\left(\bar{x}'_{2c}\right)}} \bar{x}'_{2\bar{c}}$$

es una media ponderada con los coeficientes de ponderación basados en los valores recíprocos de las varianzas. Sustituyendo los valores de W y $1 - W$ en $V\left(\bar{x}_2\right)$, calculamos el valor de la varianza mínima para el estimador de la media en segunda ocasión. Tenemos

$$V\left(\bar{x}_2\right) = \frac{V^2\left(\bar{x}_{2\bar{c}}\right)}{\left(V\left(\bar{x}'_{2c}\right) + V\left(\bar{x}_{2\bar{c}}\right)\right)^2} V^2\left(\bar{x}'_{2c}\right) + \frac{V^2\left(\bar{x}'_{2c}\right)}{\left(V\left(\bar{x}'_{2c}\right) + V\left(\bar{x}_{2\bar{c}}\right)\right)^2} V^2\left(\bar{x}_{2\bar{c}}\right) = \frac{V\left(\bar{x}_{2\bar{c}}\right)V\left(\bar{x}'_{2c}\right)}{V\left(\bar{x}_{2\bar{c}}\right) V\left(\bar{x}'_{2c}\right)}$$

y como $V\left(\bar{x}'_{2c}\right) = S^2\left(\frac{1 - \rho_{12}^2}{c} + \frac{\rho_{12}^2}{n}\right)$ y $V\left(\bar{x}_{2\bar{c}}\right) = \frac{S^2}{n - c} = \frac{S^2}{\bar{c}}$ tenemos:

$$V(\bar{x}_2) = \frac{S^2 \cdot \left(\dfrac{\left(1-\rho_{12}^2\right)n + c\rho_{12}^2}{cn} \right) \cdot \dfrac{S^2}{\bar{c}}}{S^2 \cdot \left(\dfrac{\left(1-\rho_{12}^2\right)n + c\rho_{12}^2}{cn} \right) + \dfrac{S^2}{\bar{c}}} = \frac{\left(1-\rho_{12}^2\right)\cdot n + c\rho_{12}^2}{\left(1-\rho_{12}^2\right)\cdot n + c\rho_{12}^2 + \dfrac{cn}{\bar{c}}} \cdot \frac{S^2}{\bar{c}}$$

$$= \frac{S^2}{\bar{c}} \cdot \frac{n - \rho_{12}^2(n-c)}{n - \rho_{12}^2(n-c) + \dfrac{cn}{\bar{c}}} = \frac{S^2 \cdot \left(n - \rho_{12}^2(n-c)\right)}{\bar{c}n - \rho_{12}^2\bar{c}^2 + cn} = \frac{S^2 \cdot \left(n - \rho_{12}^2(n-1)\right)}{n^2 - \rho_{12}^2\bar{c}^2}$$

Por lo tanto, ya tenemos el valor de la varianza mínima para el estimador lineal de mínima varianza de la media en segunda ocasión:

$$\boxed{V(\bar{x}_2) = S^2 \frac{n - \rho_{12}^2\bar{c}^2}{n^2 - \rho_{12}^2\bar{c}^2}}$$

En particular , $\bar{c} = 0 \Rightarrow V(\bar{x}_2) = \dfrac{S^2}{n}$ y $\bar{c} = n \Rightarrow V(\bar{x}_2) = \dfrac{S^2 \cdot n \cdot \left(1-\rho_{12}^2\right)}{n^2\left(1-\rho_{12}^2\right)} = \dfrac{S^2}{n}$

Luego podemos decir que para estimar el valor actual de $\overline{X}_2$ se obtiene la misma precisión manteniendo la muestra que cambiándola por completo en cada ocasión.

PROBLEMAS RESUELTOS

9.1. Se trata de estudiar las casas en alquiler en una población. Para ello se extrae una muestra aleatoria simple extensa y barata de tamaño 374 de las casas de un distrito y se halla que 272 casas estaban ocupadas por familias de raza blanca y 82 por otras razas. Se extrae una segunda muestra de aproximadamente una de cada cuatro casas y se obtienen los siguientes resultados respecto de la proporción de casas en alquiler:

	En alquiler	Total
Blancos	31	74
Otras razas	4	18

Estimar la proporción de casas en alquiler en la población y su error de muestreo.

Se trata de un problema de muestreo bifásico en el que la muestra de primera fase tiene de tamaño $n' = 374$ distribuyéndose entre los dos estratos con $n_1' = 272$ y $n_2' = 82$.

En segunda fase tenemos los siguientes datos por estratos:

Estrato I → Raza blanca $n_1 = 74$ $\hat{W}_1 = 272/374$ $\hat{P}_1 = 31/74$

Estrato II → Otras razas $n_2 = 18$ $\hat{W}_2 = 82/374$ $\hat{P}_2 = 4/18$

n=92

Tenemos entonces $\hat{P} = \sum_{h=1}^{2} \hat{W}_h \hat{P}_h = \frac{272}{374} \cdot \frac{31}{74} + \frac{82}{374} \cdot \frac{4}{18} = 0{,}376$.

Para hallar el error de muestreo calculamos la estimación de la varianza de la proporción a partir de la fórmula aproximada:

$$\hat{V}\left(\hat{P}\right) = \frac{n'}{n'-1}\left[\sum_h \frac{\hat{P}_h \hat{Q}_h}{n_h - 1}\left(\hat{W}_h^2 - \frac{\hat{W}_h}{n'}\right) + \frac{1}{n'}\sum_h \hat{W}_h \left(\hat{P}_n - \hat{P}\right)^2\right] =$$

$$\frac{374}{373}\left[\frac{\frac{31}{74}\cdot\frac{43}{74}}{73}\left(\left(\frac{272}{374}\right)^2 - \frac{272}{374}\right) + \frac{\frac{4}{18}\cdot\frac{14}{18}}{17}\left(\left(\frac{82}{374}\right)^2 - \frac{82}{374}\right)\right] +$$

$$\frac{1}{374}\left[\left(\frac{272}{374}\right)\left(\frac{31}{74} - 0{,}376\right)^2 + \left(\frac{82}{374}\right)\left(\frac{4}{18} - 0{,}376\right)^2\right] \cong 0{,}0025$$

El error relativo de muestreo será $\frac{\sqrt{0{,}0025}}{0{,}375} = 0{,}133 \quad (13{,}3\%)$.

9.2. Se trata de estimar una proporción a través de una encuesta para la que se dispone de un presupuesto de 300000 unidades monetarias utilizando muestreo bifásico con estratificación. La encuesta principal cuesta 1000 unidades monetarias por unidad de muestreo y se dispone de información adicional en registros a un coste de 25 unidades monetarias por unidad de muestreo que permite clasificar las unidades en dos estratos de tamaños casi iguales. Sabiendo que la proporción verdadera es 0,2 en el primer estrato y 0,8 segundo estrato, se quiere estimar los tamaños de las muestras en ambas fases n y n' óptimos y el correspondiente valor de la varianza del estimador de la proporción. Cuantificar la ganancia en precisión respecto del muestreo aleatorio simple.

Hallaremos los tamaños óptimos n' y n correspondientes a un coste total dado tales que $V\left(\hat{P}\right)$ sea mínima, escribiendo la función de Lagrange:

$$\phi = \frac{1}{n}A + \frac{1}{n'}B + \lambda(c'n' + cn - C) \text{ con } A = \left(\sum_h W_h\sqrt{P_hQ_h}\right)^2 \text{ y } B = \sum_h W_h\left(P_h - P\right)^2$$

Derivando respecto de n y n' y λ se tiene:

$$\left.\begin{aligned} \frac{\partial\phi}{\partial n} &= -\frac{A}{n^2} + \lambda c = 0 \Rightarrow \lambda = \frac{A}{cn^2} \\ \frac{\partial\phi}{\partial n'} &= -\frac{B}{n'^2} + \lambda c' = 0 \Rightarrow \lambda = \frac{B}{c'n'^2} \\ \frac{\partial\phi}{\partial\lambda} &= c'n' + cn - C = 0 \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} n &= \frac{C\sqrt{A}}{\sqrt{c}\left(\sqrt{Ac} + \sqrt{Bc'}\right)} \\ n' &= \frac{C\sqrt{B}}{\sqrt{c'}\left(\sqrt{Ac} + \sqrt{Bc'}\right)} \\ V_{ópt.}\left(\hat{\bar{X}}\right) &= \frac{\left(\sqrt{Ac} + \sqrt{Bc'}\right)^2}{C} \end{aligned}\right.$$

Tenemos como datos que $C = 300000$, $c = 1000$, $c' = 25$, $P_1 = Q_2 = 0{,}2$, $Q_1 = P_2 = 0{,}8$, $W_1 = W_2 = 0{,}5$ y $P = \sum_{h=1}^{2} W_hP_h = 0{,}5(0{,}2 + 0{,}8) = 0{,}5$. Ya podemos calcular:

$$A = \left(\sum_h W_h\sqrt{P_hQ_h}\right)^2 = \left(0{,}5\sqrt{0{,}2\cdot 0{,}8} + 0{,}5\sqrt{0{,}8\cdot 0{,}2}\right)^2 = 0{,}16$$

$$B = \sum_h W_h\left(P_h - P\right)^2 = 0{,}5\cdot(0{,}2 - 0{,}5)^2 + 0{,}5\cdot(0{,}8 - 0{,}5)^2 = 0{,}09$$

y tenemos:

$$n = \frac{C\sqrt{A}}{\sqrt{c}\left(\sqrt{Ac} + \sqrt{Bc'}\right)} = \frac{300000\sqrt{0{,}16}}{\sqrt{1000}\left(\sqrt{0{,}16\cdot 1000} + \sqrt{0{,}09\cdot 25}\right)} = 268$$

$$n' = \frac{C\sqrt{B}}{\sqrt{c'}\left(\sqrt{Ac} + \sqrt{Bc'}\right)} = \frac{300000\sqrt{0{,}09}}{\sqrt{25}\left(\sqrt{0{,}16\cdot 1000} + \sqrt{0{,}09\cdot 25}\right)} = 1272$$

$$V_{ópt.}\left(\hat{\bar{X}}\right) = \frac{\left(\sqrt{Ac} + \sqrt{Bc'}\right)^2}{C} = \frac{\left(\sqrt{0{,}16\cdot 1000} + \sqrt{0{,}09\cdot 25}\right)^2}{300000} = 0{,}0006673$$

En muestreo aleatorio simple la varianza de la proporción, considerando reposición (no olvidemos que para poblaciones grandes en muestreo bifásico pueden aproximarse todas las fórmulas por su expresión para reposición en las dos fases) será la siguiente:

$$V(\hat{P}) = \frac{PQ}{n} = \frac{0{,}5(1-0{,}5)}{300000/1000} = 0{,}0008333$$

Se observa que hay ganancia en precisión al utilizar muestreo bifásico cuantificada por (0,0008333/0,0006673-1) = 0,248, esto es, el 24,8%.

9.3. Consideremos un proceso de muestreo bifásico con estratificación. Supongamos que en la primera fase se extrae una muestra de tamaño n' = 400, y que en la segunda fase se ha tomado, una vez formados tres estratos, $n_1 = 20$, $n_2 = 10$ y $n_3 = 10$. Se conocen los siguientes resultados:

$\hat{W}_h$	$\bar{x}_h$	$\hat{S}_h^2$
0,55	2,8	15
0,32	8,2	200
0,13	26	1000

Obtener una estimación del error relativo de muestreo del estimador de la media así como una estimación de la media por intervalos al 95% de confianza.

Se considera que para poblaciones grandes, en muestreo bifásico pueden aproximarse todas las fórmulas por su expresión para reposición en las dos fases. Para estimar la varianza del estimador de la media tenemos:

$$\hat{V}\left(\hat{\bar{X}}\right) = \frac{n'}{n'-1}\left[\sum_h \frac{s_h^2}{n_h}\left(\hat{W}_h^2 - \frac{\hat{W}_h}{n'}\right) + \frac{1}{n'}\sum_h \hat{W}_h\left(\bar{x}_n - \bar{X}\right)^2\right] = \frac{400}{400-1}\left[\frac{15}{20}\left(0{,}55^2 - \frac{0{,}55}{400}\right)\right.$$

$$+ \frac{200}{10}\left(0{,}32^2 - \frac{0{,}32}{400}\right) + \frac{1000}{10}\left(0{,}13^2 - \frac{0{,}13}{400}\right) + \frac{1}{400}\left(0{,}55(2{,}8-7{,}54)^2 + 0{,}32(8{,}2-7{,}54)^2\right.$$

$$\left.\left. + 0{,}13(26-7{,}54)^2\right)\right] = 3{,}96$$

$$\hat{\bar{X}} = \sum_{h=1}^{3} \hat{W}_h \bar{x}_h = 0{,}55 \cdot 2{,}8 + 0{,}32 \cdot 8{,}2 + 0{,}13 \cdot 26 = 7{,}544$$

El error relativo será $\hat{C}v(\hat{\bar{X}}) = \dfrac{\sqrt{\hat{V}(\hat{\bar{X}})}}{\hat{\bar{X}}} = \dfrac{\sqrt{3{,}96}}{7{,}544} = 0{,}264 \quad (26{,}4\%)$

Un límite para el error de estimación al 95% vendrá dado por la anchura del intervalo de confianza, que vale $1{,}96\sqrt{3{,}96} = 3{,}9$.

Hemos visto en este capítulo que para valores grandes de n' (caso habitual) el estimador de la varianza del estimador de la media puede aproximarse por la fórmula correspondiente al estimador de la varianza del estimador de la media en muestreo estratificado en una sola fase (seguimos suponiendo reposición) sustituyendo W_h por su estimación. En nuestro caso tendríamos:

$$\hat{V}\left(\hat{\bar{X}}\right) = \sum_h \hat{W}_h^2 \frac{\hat{S}_h^2}{n_h} = \left[0{,}55^2 \frac{15}{20} + 0{,}32^2 \frac{200}{10} + 0{,}13^2 \frac{1000}{10}\right] = 4{,}12$$

El error relativo será $\hat{C}v(\hat{\bar{X}}) = \dfrac{\sqrt{\hat{V}(\hat{\bar{X}})}}{\hat{\bar{X}}} = \dfrac{\sqrt{4{,}12}}{7{,}544} = 0{,}269 \quad (26{,}9\%)$

Observamos que la pérdida en precisión es mínima por haber utilizado la aproximación citada.

9.4. Consideremos dos características X e Y medidas sobre los elementos de una población para las que conocemos los datos $\sigma_x = 2$ $\sigma_y = 4$ $\sigma_{xy} = 10$ y $\bar{X} = 10$. Se lleva a cabo un muestreo bifásico obteniendo en primera fase una muestra de tamaño $n' = 100$ con $\bar{y}' = 40{,}6$. En la segunda fase $n = 25$, $\bar{x} = 9{,}8$ e $\bar{y} = 40{,}1$. Se trata de estimar la media poblacional utilizando muestreo bifásico por regresión óptimo calculando el error relativo de muestreo y el coste total para $c' = 0$ y $c = 600$

Se tiene $\rho = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} = \dfrac{6}{2 \cdot 4} = \dfrac{6}{8} = 0{,}75$ y $b = \dfrac{\sigma_{xy}}{\sigma_y^{\,2}} = \dfrac{6}{4^2} = \dfrac{6}{16}$

El estimador por regresión para la media en el muestreo doble se halla mediante:

$$\hat{\bar{X}}_{rg} = \bar{x} + b(\bar{y}' - \bar{y}) = 9{,}8 + \frac{6}{16}(40{,}6 - 40{,}1) = 9{,}998$$

La varianza del estimador óptimo de la media se calcula mediante la expresión:

$$V\left(\hat{\bar{X}}_{rg}\right) = \frac{\left(1-\rho^2\right)\sigma_x^2}{n} + \frac{\rho^2\sigma_x^2}{n'} = \frac{\left(1-0{,}75^2\right)2^2}{25} + \frac{0{,}75^2 \cdot 2^2}{100} = 0{,}0955$$

El error relativo será $\hat{C}v(\hat{\bar{X}}_{rg}) = \dfrac{\sqrt{\hat{V}(\hat{\bar{X}}_{rg})}}{\hat{\bar{X}}_{rg}} = \dfrac{\sqrt{0{,}0955}}{9{,}998} = 0{,}0309 \quad (3{,}09\%)$

El coste total será $C = cn + c'n' = 600(25)+10(100) = 16000$.

9.5. Se utiliza una muestra aleatoria simple de tamaño 60 extraída de una población sin reposición y probabilidades iguales, para repetir una encuesta sobre sus elementos en dos ocasiones distintas. Se supone que no existe falta de respuesta y que los resultados obtenidos son los que representa la tabla adjunta. Además, se sabe que $\sigma^2 = 20$, $\rho = 0{,}7$ y $\pi = 0{,}6$.

Primera ocasión	Segunda ocasión
$\bar{x}'' = 150$	$\bar{y}'' = 160$
$\bar{x}' = 152$	$\bar{y}' = 158$

1) Hallar la estimación de cambio $\bar{y} - \bar{x}$ y su error de muestreo.
2) Hallar la estimación del cambio de mínima varianza y su error de muestreo.
3) Hallar la estimación de la media en segunda ocasión $\bar{y}$ y su error de muestreo.
4) Hallar la estimación de la media en segunda ocasión de mínima varianza y su error.

El número c de unidades muestrales comunes en las dos ocasiones se puede calcular a partir de la proporción de unidades muestrales comunes π_c y del tamaño muestral total n.

$$\pi_c = \frac{c}{n} \Rightarrow c = \pi_c \cdot n = 0{,}6 \cdot 60 = 36$$

$$\bar{x} = \frac{n-c}{n}\bar{x}'' + \frac{c}{n}\bar{x}' = \frac{60-36}{60}150 + \frac{36}{60}152 = 0{,}4 \cdot 150 + 0{,}6 \cdot 152 = 151{,}2$$

$$\bar{y} = \frac{n-c}{n}\bar{y}'' + \frac{c}{n}\bar{y}' = \frac{60-36}{60}160 + \frac{36}{60}158 = 0{,}4 \cdot 160 + 0{,}6 \cdot 158 = 158{,}8$$

Para la *estimación del cambio y su error* tenemos entonces:

$$\hat{\partial} = \bar{y} - \bar{x} = 158{,}8 - 151{,}2 = 7{,}6$$

$$V\left(\hat{\partial}\right) = 2\frac{S^2}{n}\left[1 - \rho_{12}\pi_c\right] \cong 2\frac{20}{60}\left[1 - 0{,}7 \cdot 0{,}6\right] = 0{,}38666$$

El *estimador del cambio de mínima varianza y su error* vienen dados por:

$$\hat{\Delta} = W\left(\bar{y}' - \bar{x}'\right) + \left(1 - W\right) \cdot \left(\bar{y}'' - \bar{x}''\right) \text{ con } W = \frac{\pi_c}{1 - \rho_{12}\left(1 - \pi_c\right)} = \frac{0{,}6}{1 - 0{,}7 \cdot 0{,}4} = 0{,}8333$$

luego ya tenemos $\hat{\Delta} = 0{,}8333\left(158 - 152\right) + \left(1 - 0{,}8333\right) \cdot \left(160 - 150\right) = 6{,}66666$

$$V\left(\hat{\Delta}\right) = \frac{2S^2\left(1 - \rho_{12}\right)}{\left[1 - \rho_{12}\left(1 - \pi_c\right)\right] \cdot n} \cong \frac{2 \cdot 20\left(1 - 0{,}7\right)}{\left[1 - 0{,}7\left(1 - 0{,}6\right)\right] \cdot 60} = 0{,}277$$

El *estimador de la media en segunda ocasión y su error* se calculan como:

$$\bar{y} = \frac{n-c}{n}\bar{y}'' + \frac{c}{n}\bar{y}' = \frac{60-36}{60}160 + \frac{36}{60}158 = 0{,}4 \cdot 160 + 0{,}6 \cdot 158 = 158{,}8$$

$$V\left(\bar{y}\right) = \frac{S^2}{n} \cong \frac{20}{60} = 0{,}333$$

Utilizaremos el estimador *estimador lineal de mínima varianza de la media para la segunda ocasión* combinado definido por:

$$\bar{y} = W\left[\bar{y}' + \rho\left(\bar{x} - \bar{x}'\right)\right] + \left(1 - W\right)\bar{y}'' = 0{,}65\left[158 + 0{,}7\left(151{,}2 - 152\right)\right] + \left(1 - 0{,}65\right)160 = 159$$

Los cálculos necesarios son los siguientes:

$$W = \frac{V\left(\bar{x}_{2\bar{c}}\right)}{V\left(\bar{x}'_{2c}\right) + V\left(\bar{x}_{2\bar{c}}\right)} = \frac{0{,}833}{0{,}446 + 0{,}833} = 0{,}65$$

$$V\left(\bar{x}'_{2c}\right) = S^2\left(\frac{1 - \rho_{12}^2}{c} + \frac{\rho_{12}^2}{n}\right) = 20\left(\frac{1 - 0{,}7^2}{32} + \frac{0{,}7^2}{60}\right) = 0{,}446 \quad V\left(\bar{x}_{2\bar{c}}\right) = \frac{S^2}{n-c} = \frac{20}{60-36} = 0{,}833$$

El error de muestreo del estimador de varianza mínima viene dado por:

$$V(\bar{y}) = \frac{S^2 \cdot \left(n - \rho_{12}^2 (n-1)\right)}{n^2 - \rho_{12}^2 \bar{c}^2} = \frac{20 \cdot \left(60 - 0{,}7^2 (60-1)\right)}{60^2 - 0{,}7^2 (60-36)^2} = 0{,}29$$

9.6. Se utiliza una muestra aleatoria simple de tamaño 100 de una población de 1000 personas sin reposición y probabilidades iguales para repetir una encuesta sobre sus elementos en dos ocasiones sucesivas preguntando sobre un carácter dicotómico. Se obtienen los resultados de la tabla adjunta.

$O_1 \rightarrow$ $O_2 \downarrow$	*Sí*	*No*	*Total*
Sí	80	5	85
No	10	5	15
Total	90	10	100

Hallar ρ y calcular el error de muestreo del estimador diferencia de proporciones con contestación afirmativa entre la segunda y la primera ocasión.

$$\hat{D} = \hat{P}_2 - \hat{P}_1 \Rightarrow \hat{V}(\hat{D}) = \hat{V}(\hat{P}_2) + \hat{V}(\hat{P}_1) - 2Cov(\hat{P}_1, \hat{P}_2) = (1-f)\frac{\hat{P}_2(1-\hat{P}_2)}{n-1} +$$

$$(1-f)\frac{\hat{P}_1(1-\hat{P}_1)}{n-1} - 2(1-f)\frac{\sum_{i=1}^{n} X_{1i} \cdot X_{2i} - n\hat{P}_1\hat{P}_2}{n(n-1)} = \left(1 - \frac{10}{100}\right)\frac{\frac{85}{100}(1-\frac{85}{100})}{100-1} +$$

$$\left(1 - \frac{10}{100}\right)\frac{\frac{90}{100}(1-\frac{90}{100})}{100-1} + 2\left(1 - \frac{10}{100}\right)\frac{80 - 100\frac{90}{100}\frac{85}{100}}{n(n-1)} = 0{,}00134$$

Con los datos de la tabla se comprueba fácilmente que $\sum_{i=1}^{n} X_{1i} \cdot X_{2i} = 80$.

El coeficiente de correlación se calculará de la siguiente forma:

$$\rho = \frac{Cov(\hat{P}_1, \hat{P}_2)}{\sqrt{\hat{V}(\hat{P}_1)}\sqrt{\hat{V}(\hat{P}_2)}} = \frac{0{,}00032}{\sqrt{0{,}00082}\sqrt{0{,}00116}} = 0{,}3$$

EJERCICIOS PROPUESTOS

9.1. Se destinan 3000 unidades monetarias a una encuesta para estimar una proporción. La encuesta principal costará 10 unidades monetarias por unidad de muestreo. Se dispone de información en registros, a un coste de 0,25 unidades monetarias por unidad de muestreo, que permite la clasificación de las unidades en dos estratos de tamaños casi iguales. Si la proporción verdadera es 0,2 en el estrato 1 y 0,8 en el estrato 2, estimar n y n' óptimas y el valor resultante de $V(p_{st})$. ¿Produce el muestreo bifásico alguna ganancia en precisión sobre el muestreo aleatorio simple?

9.2. Si $\rho = 0{,}8$ en muestreo doble para regresión, ¿cómo debe ser n' con relación a n, si la pérdida en precisión debida a errores de muestreo en la media de la muestra grande se desea que sea menor del 10%?

9.3. En una aplicación de muestreo bifásico por regresión la muestra pequeña es de tamaño 87 y la grande de tamaño 300. Para la muestra pequeña conocemos los siguientes datos:

$$\sum_i (X_i - \bar{x})^2 = 17283 \qquad \sum_i (X_i - \bar{x})(Y_i - \bar{y}) = 5114 \qquad \sum_i (Y_i - \bar{y})^2 = 3248$$

Calcular el error estándar de la estimación de la regresión de $\bar{X}$.

9.4. En un muestreo en dos ocasiones se supone que $S_1=S_2=S$ y que las muestras son grandes de modo que los coeficientes de regresión de X_{2i} respecto de X_{1i} y de X_{1i} respecto de X_{2i} en la parte apareada de las muestras en las dos ocasiones son ambas efectivamente iguales a ρ. Demostrar que si las estimaciones $\bar{x}_1$ y $\bar{x}_2$ se construyen usando la regresión de X_{1i} respecto de X_{2i} se tiene:

$$v(\bar{x}_2 - \bar{x}_1) = \frac{2S^2(1-\rho)}{(n-\rho u)} \qquad v(\bar{x}_2 + \bar{x}_1) = \frac{2S^2(1+\rho)}{(n+\rho u)}$$

CAPÍTULO

10

MUESTREO ESTADÍSTICO MEDIANTE SPSS

OBJETIVOS

1. Presentar métodos automatizados de tratamiento del muestreo estadístico.
2. Analizar las posibilidades en muestreo del software SPSS a partir de la versión 12.
3. Utilizar el asistente de muestreo de SPSS para la creación de planes de muestreo en diseños complejos.
4. Utilizar el asistente de muestreo de SPSS para la modificación y ejecución de planes de muestreo en diseños complejos.
5. Estudiar las posibilidades del asistente de preparación de análisis de SPSS para la creación de un plan de análisis en una muestra compleja.
6. Estudiar las posibilidades del asistente de preparación de análisis de SPSS para la modificación de un plan de análisis en una muestra compleja.
7. Realizar cálculos en muestra complejas con SPSS.
8. Obtener frecuencias, descriptivos, tablas de contingencia, razones y otros estimadores y sus errores en muestras complejas con SPSS.

ÍNDICE

1. SPSS y el muestreo estadístico.

2. Diseños complejos y el asistente de muestreo. Creación de un nuevo plan de muestreo.

3. Asistente de muestreo: Modificar un plan existente.

4. Asistente de muestreo: ejecutar un plan de muestreo dado.

5. Preparación de una muestra compleja para su análisis: Creación de un nuevo plan de análisis.

6. Preparación de una muestra compleja para su análisis: Modificar un plan de análisis existente.

7. Cálculos en muestras complejas: frecuencias, descriptivos, tablas de contingencia y razones.

SPSS Y EL MUESTREO ESTADÍSTICO

Un supuesto inherente a los procedimientos de análisis en los paquetes de software tradicionales es que las observaciones del archivo de datos de trabajo representan una muestra aleatoria simple de la población de interés. Este supuesto es insostenible para un número cada vez mayor de empresas e investigadores que consideran más económico y cómodo obtener las muestras de una forma más estructurada. La opción *Muestras complejas de SPSS* (opción presente en el programa a partir de la versión 12) permite seleccionar una muestra de acuerdo con un diseño complejo e incorporar las especificaciones del diseño al análisis de los datos para asegurar la validez de los resultados.

En SPSS, una muestra compleja puede ser distinta de una muestra aleatoria simple en muchos aspectos. En una muestra aleatoria simple, las unidades de muestreo individuales se seleccionan aleatoriamente con la misma probabilidad y sin reposición (SR) directamente a partir de la totalidad de la población. Por el contrario, una muestra compleja determinada puede tener en SPSS alguna o todas las características siguientes:

Estratificación: El muestreo estratificado implica seleccionar muestras independientemente dentro de los subgrupos de la población que no se solapen o estratos. Por ejemplo, los estratos pueden ser grupos socioeconómicos, categorías laborales, grupos de edad o grupos étnicos. Con la estratificación, puede asegurar que los tamaños muestrales de los subgrupos de interés son adecuados, mejorar la precisión de las estimaciones globales y utilizar distintos métodos de muestreo entre los diferentes estratos.

Conglomerados: El muestreo por conglomerados implica la selección de grupos de unidades muestrales o conglomerados. Por ejemplo, los conglomerados pueden ser escuelas, hospitales o zonas geográficas y las unidades muestrales pueden ser alumnos, pacientes o ciudadanos. El conglomerado es común en los diseños polietápicos y en las muestras de zona (geográfica).

Múltiples etapas: En el muestreo polietápico, se selecciona una muestra de primera etapa basada en conglomerados. A continuación, se crea una muestra de segunda etapa extrayendo submuestras a partir de los conglomerados seleccionados. Si la muestra de segunda etapa está basada en subconglomerados, entonces puede añadir una tercera etapa a la muestra. Por ejemplo, en la primera etapa de una encuesta, se podría extraer una muestra de ciudades. A continuación, y a partir de las ciudades seleccionadas, se podrían muestrear unidades familiares. Finalmente, a partir de las unidades familiares seleccionadas, se podría encuestar a individuos. Los Asistentes de muestreo y preparación del análisis permiten especificar tres etapas en un diseño.

Muestreo no aleatorio: Cuando es difícil obtener la muestra aleatoriamente, las unidades se pueden muestrear sistemáticamente (con un intervalo fijo) o secuencialmente.

Probabilidades de selección desiguales: Cuando se muestrean conglomerados que contienen números de unidades desiguales, puede utilizar el muestreo probabilístico proporcional al tamaño (PPS) para que la probabilidad de selección del conglomerado sea igual a la proporción de unidades que contiene. El muestreo PPS también puede utilizar esquemas de ponderación más generales para seleccionar unidades.

Muestreo no restringido: El muestreo no restringido selecciona las unidades con reposición (CR), por lo que se puede seleccionar más de una vez una unidad individual para la muestra.

Ponderaciones muestrales: Las ponderaciones muestrales se calculan automáti-camente al extraer una muestra compleja y de forma ideal se corresponden con la <<frecuencia>> que cada unidad muestral representa en la población objetivo. Por lo tanto, la suma de las ponderaciones muestrales debe estimar el tamaño de la población. Los procedimientos de análisis de muestras complejas requieren las ponderaciones muestrales para poder analizar correctamente una muestra compleja.

DISEÑOS COMPLEJOS Y EL ASISTENTE DE MUESTREO. CREACIÓN DE UN NUEVO PLAN DE MUESTREO

El *Asistente de muestreo* le guía a través de los pasos para crear, modificar o ejecutar un archivo de plan de muestreo. Antes de utilizar el Asistente, debe tener en mente una población objetivo bien definida, una lista de las unidades muestrales y un diseño muestral adecuado.

Para crear un nuevo plan de muestreo (por ejemplo, muestreo estratificado del 10% por barrios en el fichero *Venta de casas [por barrios].sav*), elija en los menús *Analizar→ Muestras complejas → Seleccionar una muestra...*(Figura 10-1). En el *Asistente de muestreo* seleccione *Diseñar una muestra* y elija un nombre de archivo de plan (PLAN1.CSPLAN) para guardar el plan de muestreo (Figura 10-2). Pulse *Siguiente* para ir al paso *Variables en el diseño* (Figura 10-3), donde puede definir estratos, conglomerados e introducir ponderaciones muestrales.

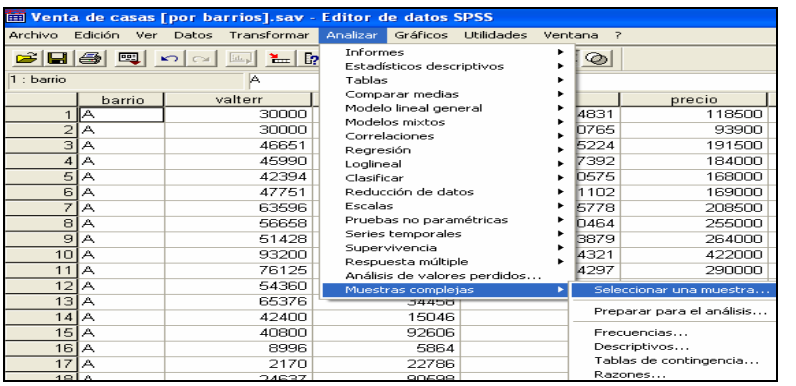


Figura 10-1

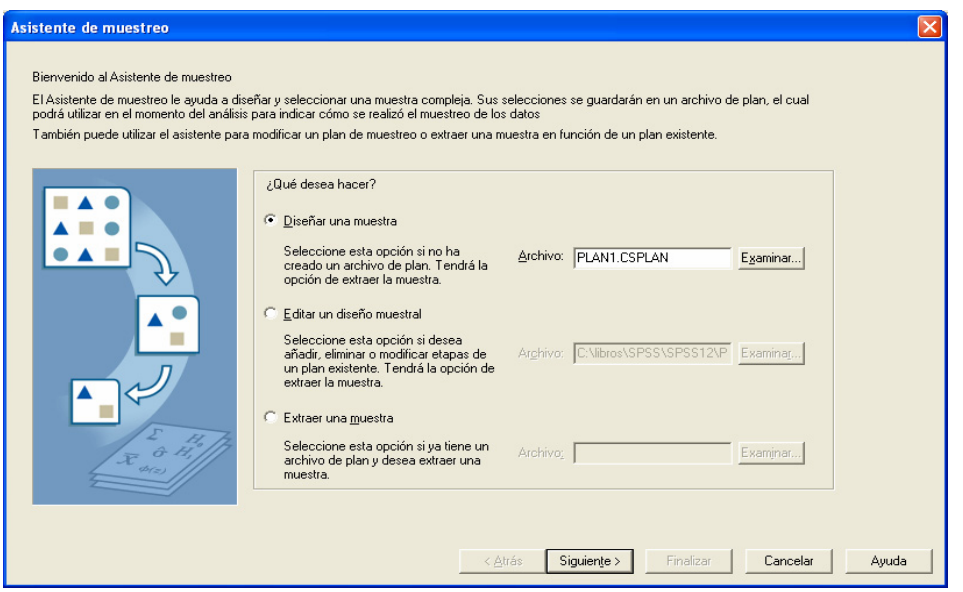


Figura 10-2

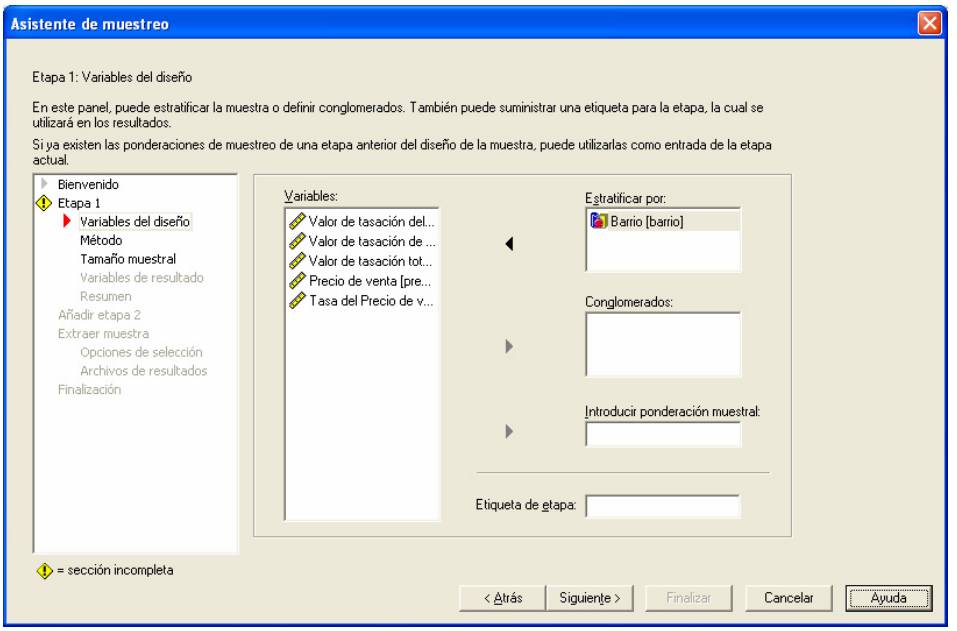


Figura 10-3

Este paso permite seleccionar las variables de estratificación y conglomeración en el campo *Variables* arrastrándolas a los campos *Estratificar por* y *Conglomerados* respectivamente, y definir ponderaciones muestrales de entrada en el campo *Introducir ponderación muestral* (si el diseño muestral actual forma parte de un diseño muestral mayor, puede disponer de ponderaciones muestrales de una etapa anterior del diseño mayor, en cuyo caso puede especificar una variable numérica que contenga estas ponderaciones en la primera etapa del diseño actual calculándose las ponderaciones muestrales automáticamente para las etapas posteriores del diseño actual). También puede especificar una etiqueta para la etapa en el campo *Etiqueta de etapa* (se utiliza en los resultados para facilitar la identificación de la información por etapas).

En la parte izquierda de cada paso del *Asistente de muestreo* se muestra un esquema de los titulares de todos los pasos. Puede navegar por el Asistente al pulsar el nombre de uno de los pasos activados en el esquema. Los pasos están activados cuando todos los pasos anteriores sean válidos, es decir, si cada uno de los pasos anteriores dispone de las especificaciones mínimas necesarias para ese paso. Consulte la ayuda de los pasos individuales para obtener más información sobre los motivos por los que un paso determinado puede no ser válido.

A continuación, para ir al paso *Método*, pulsamos en *Método* en la parte izquierda de la pantalla del Asistente para obtener la Figura 10-4, en cuyo campo *Método* elegimos el tipo de muestreo (aleatorio, sistemático, con o sin reposición, etc.).

Algunos tipos de muestreo permiten elegir entre realizar un muestreo con reposición (CR) o sin reposición (SR). Si desea obtener más información, consulte las descripciones de los tipos. Tenga en cuenta que algunos tipos de probabilidad proporcional al tamaño (PPS) están disponibles sólo cuando se han definido conglomerados y todos los tipos de PPS están disponibles sólo en la primera etapa de un diseño. Además, los métodos SR están disponibles sólo en la última etapa de un diseño.

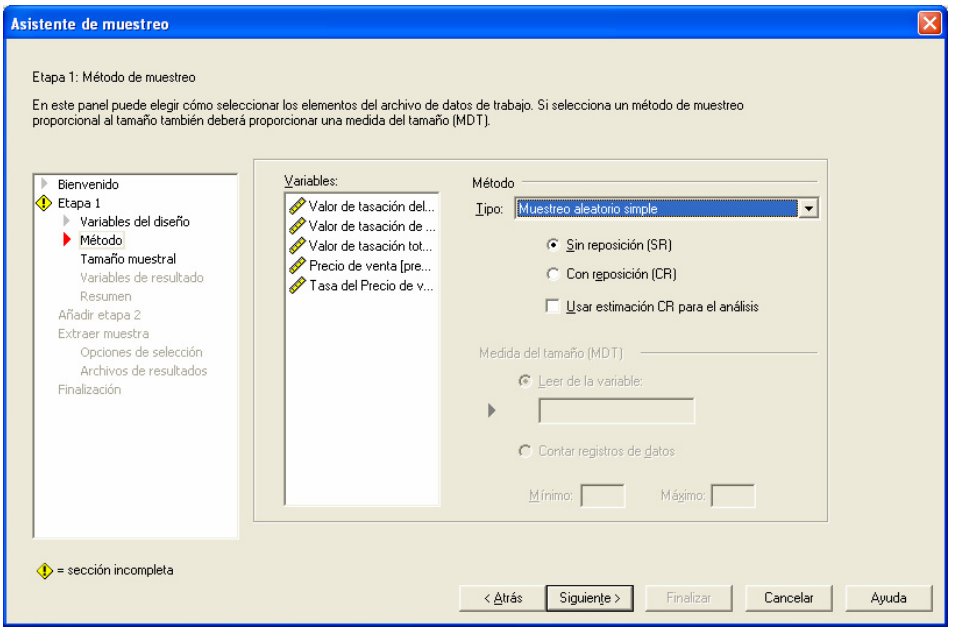


Figura 10-4

En el ***Muestreo aleatorio simple*** las unidades se seleccionan con probabilidad igual. Se pueden seleccionar con o sin reposición. En el ***Muestreo sistemático simple*** las unidades se seleccionan con un intervalo fijo en todo el marco muestral (o en los estratos, si se han especificado) y se extraen sin reposición. Se selecciona una unidad aleatoriamente dentro del primer intervalo como el punto inicial. En el ***Muestreo secuencial simple*** las unidades se seleccionan de forma secuencial con probabilidad igual y sin reposición. El ***Muestreo con probabilidad proporcional al tamaño*** es un método de primera etapa que selecciona unidades de forma aleatoria con probabilidad proporcional al tamaño. Se puede seleccionar cualquier unidad con reposición; sólo se puede realizar muestreo sin reposición de los conglomerados. El ***Muestreo sistemático proporcional al tamaño*** es un método de primera etapa que selecciona unidades de forma sistemática con probabilidad proporcional al tamaño. Se seleccionan sin reposición. El ***Muestreo secuencial proporcional al tamaño*** es un método de primera etapa que selecciona unidades de forma secuencial con probabilidad proporcional al tamaño del conglomerado y sin reposición.

El ***Muestreo de Brewer proporcional al tamaño*** es un método de primera etapa que selecciona dos conglomerados de cada estrato con probabilidad proporcional al tamaño del conglomerado y sin reposición. Se debe especificar una variable de conglomeración para utilizar este método. El ***Muestreo de Murthy proporcional al tamaño*** es un método de primera etapa que selecciona dos conglomerados de cada estrato con probabilidad proporcional al tamaño del conglomerado y sin reposición. Se debe especificar una variable de conglomeración para utilizar este método. El ***Muestreo de Sampford proporcional al tamaño*** es un método de primera etapa que selecciona más de dos conglomerados de cada estrato con probabilidad proporcional al tamaño del conglomerado y sin reposición. Es una extensión del método de Brewer. Se debe especificar una variable de conglomeración para utilizar este método. Por defecto, el método de estimación se especifica en el archivo de plan de manera coherente con el método de muestreo seleccionado, pero la opción ***Usar estimación CR para el análisis*** permite utilizar la estimación con reposición incluso si el método de muestreo implica la estimación SR. Esta opción solamente está disponible en la etapa 1. Si se selecciona un método PPS, se deberá especificar una medida del tamaño que defina el tamaño de cada unidad en el campo ***Medida del tamaño (MDT)***.

Estos tamaños pueden definirse explícitamente en una variable o se pueden calcular a partir de los datos. Opcionalmente, se pueden establecer los límites inferior y superior de la MDT, anulando cualquier valor encontrado en la variable MDT o calculado a partir de los datos. Estas opciones solamente están disponibles en la etapa 1.

A continuación, para ir al paso *Tamaño muestral*, pulsamos en *Tamaño muestral* en la parte izquierda de la pantalla del Asistente para obtener la Figura 10-5 .

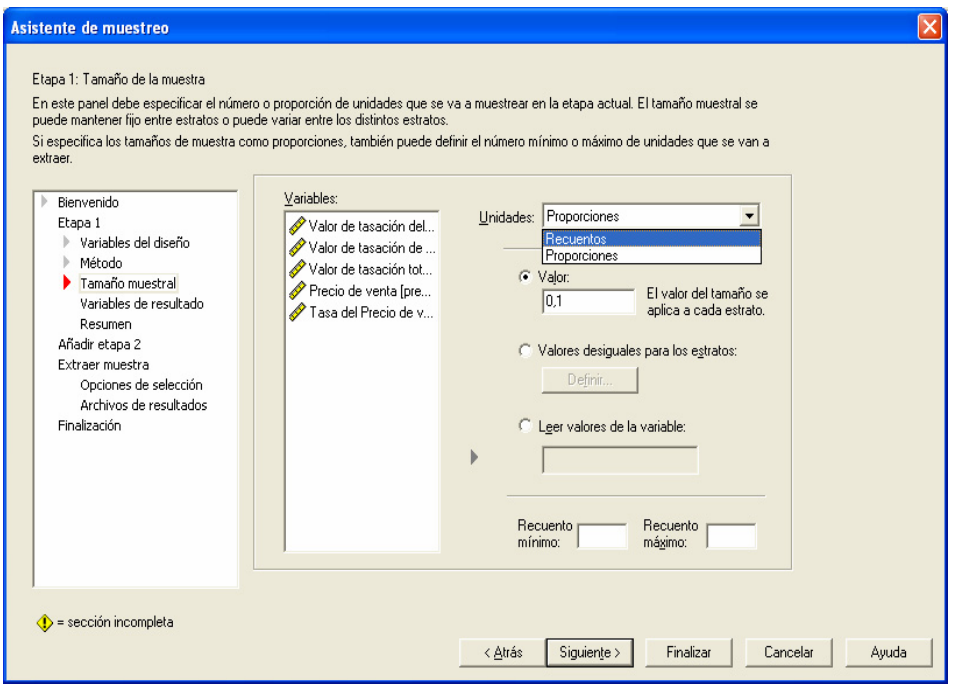


Figura 10-5

Este paso permite especificar el número o la proporción de unidades que se van a muestrear dentro de la etapa actual. El tamaño muestral puede ser fijo o variar entre estratos. Para el propósito de especificar el tamaño muestral, se pueden utilizar los conglomerados elegidos en etapas anteriores para definir estratos. En el campo *Unidades* puede especificar un tamaño muestral exacto o una proporción de unidades a muestrear. En el campo *Valor* se aplica un valor particular a todos los estratos. Si se selecciona *Recuentos* como la unidad métrica, se deberá introducir un entero positivo. Si se selecciona *Proporciones*, se deberá introducir un valor no negativo (a no ser que se realice una muestra con reposición, los valores de proporción no deberán ser mayores que 1). El campo *Valores desiguales para estratos* permite introducir distintos valores de tamaño para cada estrato a través del cuadro de diálogo *Definir tamaños desiguales*. El campo *Leer valores de la variable* permite seleccionar una variable numérica que contenga los valores de tamaño para los estratos. Si se selecciona *Proporciones*, se tiene la opción de establecer los límites inferior y superior para el número de unidades muestreadas.

A continuación, para ir al paso *Variables de resultado*, pulsamos en *Variables de resultado* en la parte izquierda de la pantalla del Asistente. Se obtiene la Figura 10-6.

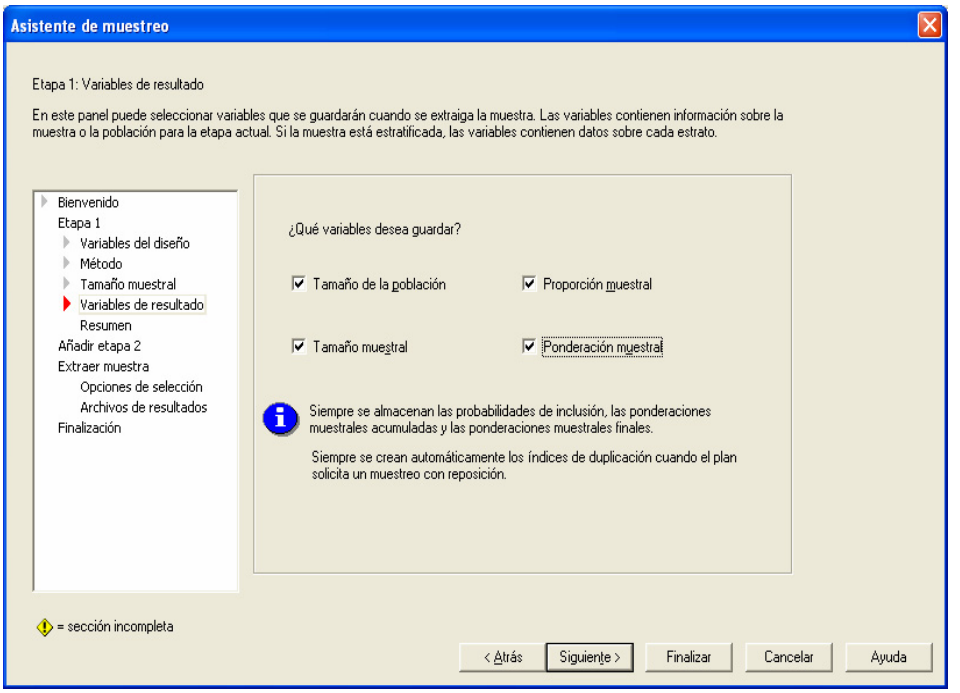


Figura 10-6

Este paso permite elegir las variables que desea guardar cuando se extraiga la muestra. *Tamaño poblacional* recoge el número estimado de unidades en la población de una etapa dada. El nombre raíz de la variable guardada es *TamañoPoblación_*. *Proporción muestral* recoge la tasa de la muestra en una etapa dada. El nombre raíz de la variable guardada es *TasaMuestreo_*. *Tamaño muestral* recoge el número de unidades extraídas en una etapa dada. El nombre raíz de la variable guardada es *TamañoMuestra_*. *Ponderación muestral* recoge la inversa de las probabilidades de inclusión.

El nombre raíz de la variable guardada es *PonderaciónMuestra_*. Algunas variables por etapa se generan automáticamente. Entre éstas se incluyen *Probabilidades de inclusión* (proporción de unidades extraídas en una etapa dada con nombre raíz de la variable guardada *ProbabilidadInclusión_*), *Ponderación acumulada* (ponderación de la muestra acumulada a lo largo de las etapas anteriores a la actual e incluyendo esta última con nombre raíz de la variable guardada), *PonderaciónMuestraAcumulada_*, *Índice* (identifica las unidades seleccionadas varias veces dentro de una etapa dada con nombre raíz de la variable guardada *Índice_*), etc.. Los nombres raíz de la variable guardada incluyen un sufijo entero que refleja el número de la etapa, por ejemplo, *TamañoPoblación*_1_ para el tamaño de la población guardada de la etapa 1.

A continuación, para ir al paso *Resumen*, pulsamos en *Resumen* en la parte izquierda de la pantalla del Asistente. Se obtiene la Figura 10-7. Se trata del último paso de cada etapa que proporciona un resumen de las especificaciones del diseño muestral hasta la etapa actual. A partir de aquí, puede pasar a la siguiente etapa (creándola si es necesario en *Añadir etapa 2*) o definir las opciones para extraer la muestra.

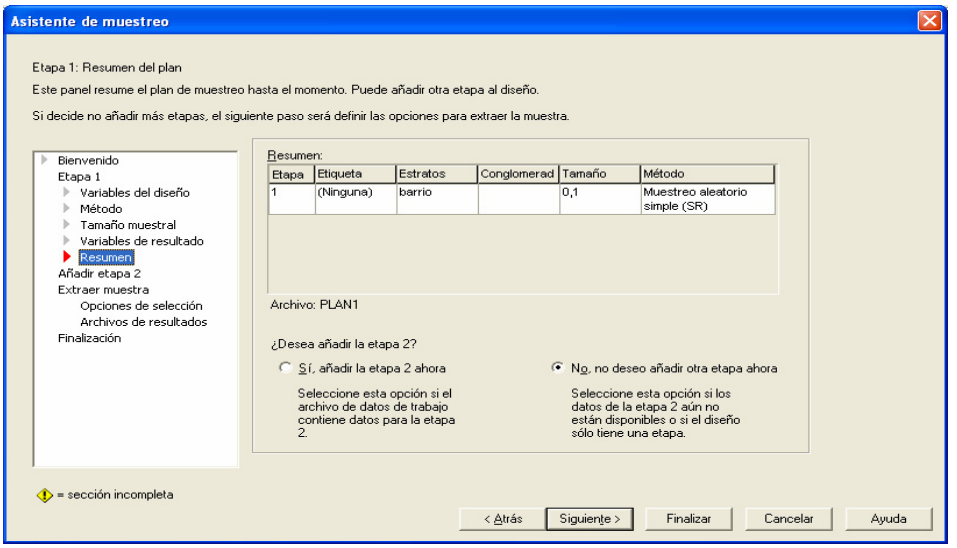


Figura 10-7

Ya estamos en condiciones de ***extraer la muestra*** según el diseño definido en los pasos anteriores. Para ello elegimos *Extraer muestra* → *Opciones de selección* en la parte izquierda de la pantalla del *Asistente de muestreo*. También puede controlar otras opciones del muestreo, como la semilla aleatoria y el tratamiento de los valores perdidos (Figura 10-8). *Extraer muestra*, además de elegir si desea extraer una muestra, también puede elegir ejecutar parte del diseño muestral. Las etapas se deben extraer en orden; es decir, la etapa 2 no se puede extraer a menos que ya se haya extraído la etapa 1. Al editar o ejecutar un plan, no puede volver a muestrear etapas bloqueadas. El campo *Semilla* permite elegir un valor de semilla para la generación de números aleatorios. El campo *Incluye los valores perdidos definidos por el usuario* determina si los valores perdidos definidos por el usuario son tratados como válidos. Si es así, los valores perdidos definidos por el usuario se tratan como una categoría diferente. El campo *Los datos ya están ordenados* permite acelerar el proceso de selección si el marco muestral está clasificado previamente por los valores de las variables de estratificación.

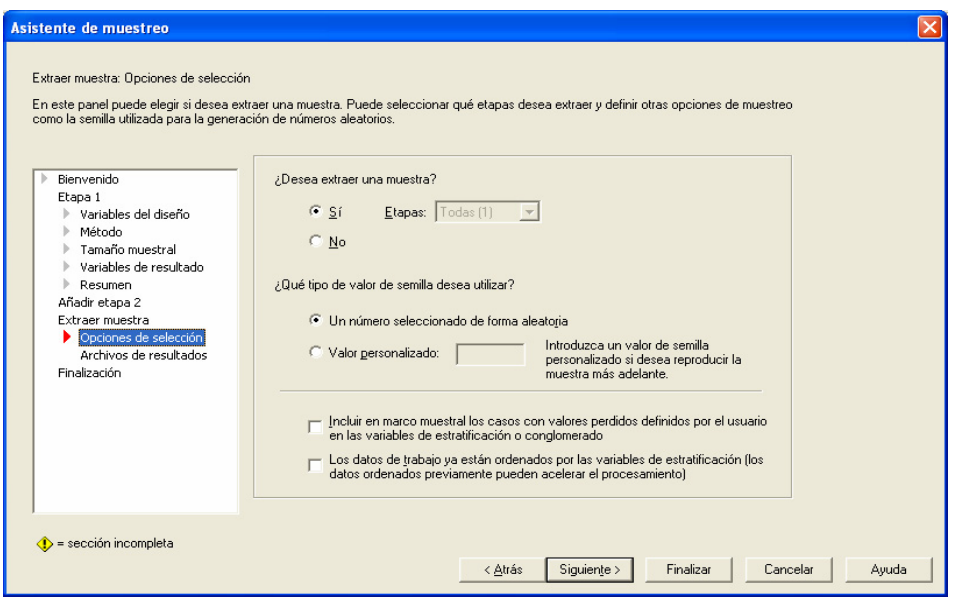


Figura 10-8

Realizado el diseño y extraída la muestra, sólo resta ***guardar los resultados*** adecuadamente. Para ello se selecciona *Extraer muestra → Archivos de resultados* en la parte izquierda de la pantalla del *Asistente de muestreo* (Figura 10-9). Este paso permite elegir dónde dirigir los casos muestreados, las variables de ponderación, las probabilidades conjuntas y las reglas de selección de casos.

Las opciones de *¿Dónde desea almacenar los datos de la muestra?* permiten determinar dónde se escribe el resultado de la muestra. Se puede añadir al archivo de datos de trabajo o guardar en un archivo externo. Si se especifica un archivo externo, se guardan en el archivo las variables de los resultados del muestreo y las variables del archivo de datos de trabajo para los casos seleccionados. Las opciones de *¿Dónde desea guardar las probabilidades conjuntas?* permiten determinar dónde se escriben las probabilidades conjuntas. Las probabilidades conjuntas se producen si se seleccionan la probabilidad proporcional al tamaño sin reposición, el muestreo de Brewer proporcional al tamaño, el muestreo de Sampford proporcional al tamaño o el método de Murthy proporcional al tamaño y la estimación con reposición no se especifica. En cuanto al campo *Guardar reglas de selección de casos*, si está construyendo la muestra por etapas, es posible que quiera guardar las reglas de selección de casos en un archivo de texto. Son útiles para construir el submarco de las etapas posteriores.

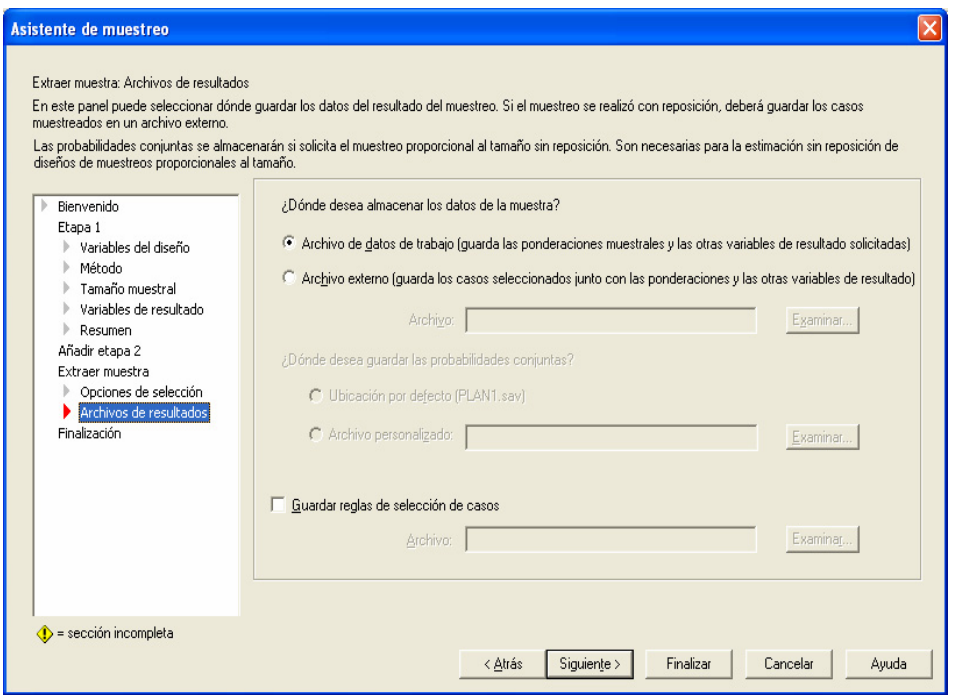


Figura 10-9

Ya sólo resta ***finalizar*** el proceso adecuadamente. Para ello se selecciona *Extraer muestra → Finalización* en la parte izquierda de la pantalla del *Asistente de muestreo* (Figura 10-10). Puede guardar el archivo de plan y extraer la muestra ahora o pegar las selecciones en una ventana de sintaxis. Al editar un plan, puede guardar el plan editado en un archivo nuevo o sobrescribir el archivo de plan existente.

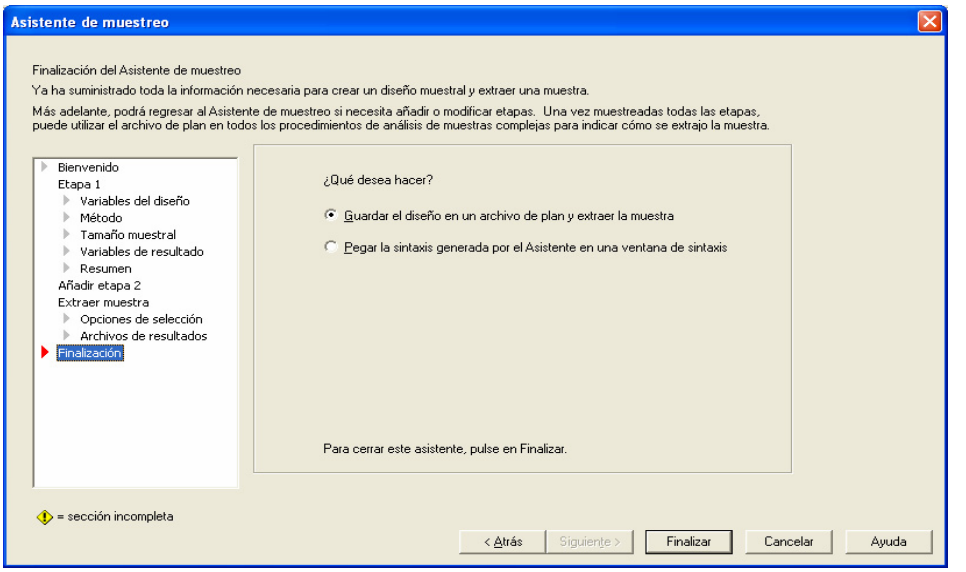


Figura 10-10

Al pulsar en *Finalizar* en la Figura 10-10 se obtiene la salida del procedimiento con la sintaxis (Figura 10-11) y un resumen para las etapas (Figura 10-12).

```
* Asistente de muestreo.
CSPLAN SAMPLE
 /PLAN FILE='PLAN1'
 /PLANVARS SAMPLEWEIGHT=PonderaciónMuestral_Final_
 /PRINT PLAN
 /DESIGN  STRATA= barrio
   /METHOD TYPE= SIMPLE_WOR  ESTIMATION= DEFAULT
   /RATE VALUE=0.1
   /STAGEVARS
     INCLPROB( ProbabilidadInclusión_1_ )
     CUMWEIGHT( PonderaciónMuestralAcumulada_1_ )
     POPSIZE( TamañoPoblación_1_ )
     SAMPSIZE( TamañoMuestral_1_ )
     RATE( TasaMuestreo_1_ )
     WEIGHT( PonderaciónMuestral1_ ).
```

Muestras complejas: Plan

Advertencia

Este procedimiento no comprueba la consistencia entre el archivo de datos de trabajo y el archivo del plan. Se recomienda estudiar la tabla de resultados o del archivo del plan para comprobar la consistencia antes de realizar la selección o el análisis.

Figura 10-11

```
CSSELECT
 /PLAN FILE='PLAN1'
 /CRITERIA STAGES = 1 SEED = RANDOM
 /CLASSMISSING EXCLUDE
 /DATA RENAMEVARS
 /PRINT SELECTION.
```

Muestras complejas: Selección

Resumen para la etapa 1

	Número de unidades muestreadas		Proporción de unidades muestreadas	
Barrio	Solicitados	Reales	Solicitados	Reales
A	4	4	10,0%	9,5%
B	32	32	10,0%	10,0%
C	26	26	10,0%	10,1%
D	47	47	10,0%	10,1%
E	50	50	10,0%	10,0%
F	37	37	10,0%	9,9%
G	48	48	10,0%	10,0%

Archivo del plan: C:\Archivos de programa\SPSS12\PLAN1

Figura 10-12

También se obtiene un resumen sobre las distintas etapas de selección de la muestra, que se presenta a continuación.

Resumen

			Etapa 1
Variables	Estratificación	1	Barrio
Información de la muestra	Método de selección		Muestreo aleatorio simple sin reposición
	Proporción de unidades muestreadas		,1
	Variables creadas o modificadas	Probabilidad de inclusión (selección) según etapa	ProbabilidadInclusión_1_
		Ponderación de muestreo acumulada según etapa	PonderaciónMuestralAcumulada_1_
		Tamaño de la población según etapa	TamañoPoblación_1_
		Tamaño de la muestra según etapa	TamañoMuestral_1_
		Tasa de muestreo según etapa	TasaMuestreo_1_
		Ponderación de muestreo según etapa	PonderaciónMuestral1_
Información sobre el análisis	Supuestos del estimador		Muestreo de probabilidad igual sin reposición
	Probabilidad de inclusión		A partir de la variable ProbabilidadInclusión_1_

Archivo del plan: C:\Archivos de programa\SPSS12\PLAN1Variable de ponderación: PonderaciónMuestral_Final_

ASISTENTE DE MUESTREO: MODIFICAR UN PLAN EXISTENTE

Para modificar un plan de muestreo existente, por ejemplo para guardar la muestra estratificada anterior en una archivo nuevo de nombre PLAN2.SAV, elija en los menús: *Analizar* → *Muestras complejas* → *Seleccionar una muestra...,* seleccione *Editar un diseño muestral* y elija el archivo de plan anterior PLAN1.CSPLAN para editar (Figura 10-13). Pulse *Siguiente* para continuar usando el *Asistente*. Revise el plan de muestreo del paso *Resumen del plan* (Figura 10-14), y a continuación pulse *Siguiente*. En *Extraer muestra* → *Archivos de resultados* especifique el archivo para guardar la muestra (Figura 10-15). Vaya al paso final y especifique un nombre nuevo para el archivo de plan editado (Figura 10-16). Si lo desea, tiene la posibilidad de *Especificar las etapas que ya se han muestreado* y *Eliminar etapas del plan*.

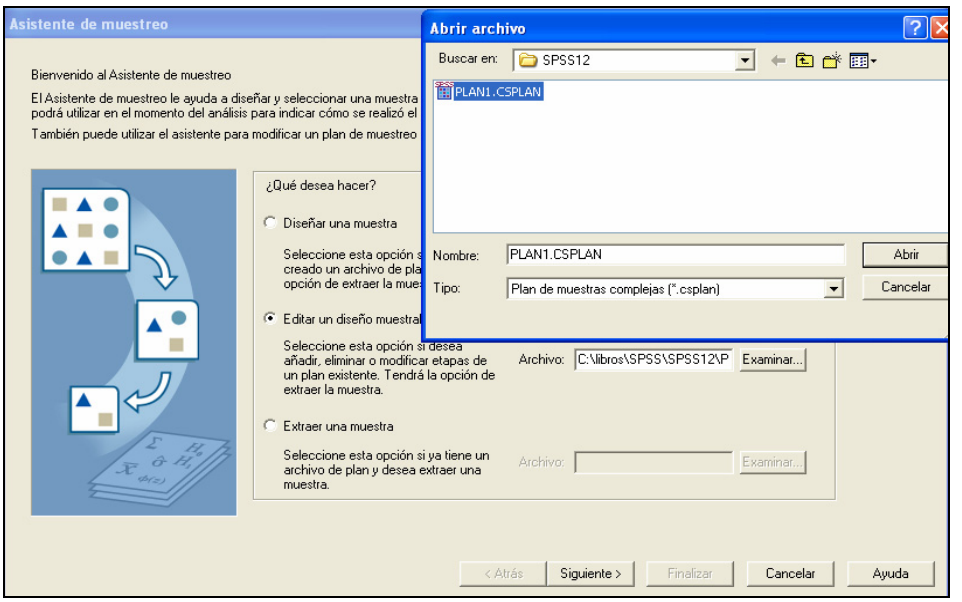


Figura 10-13

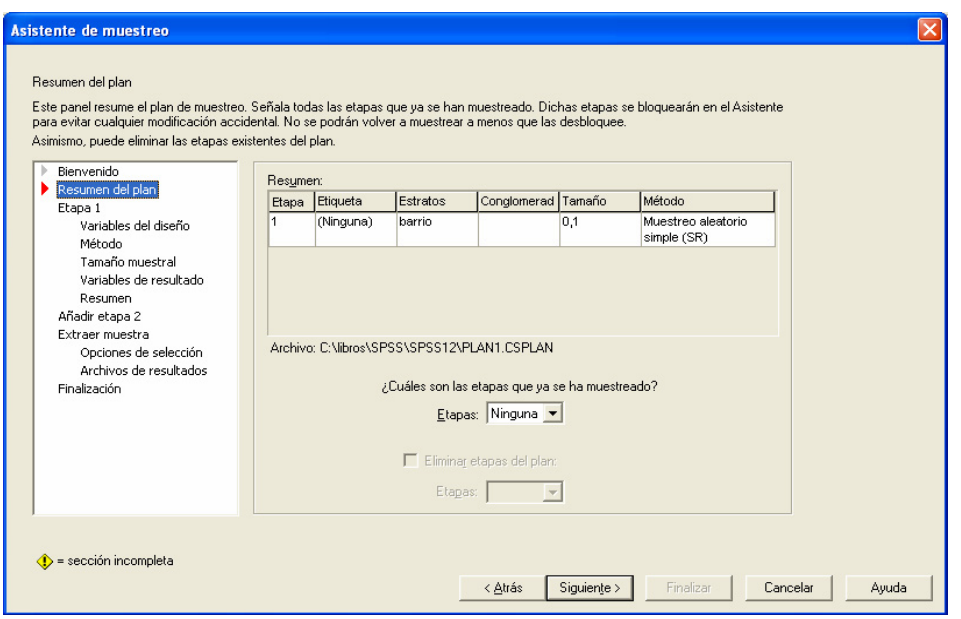


Figura 10-14

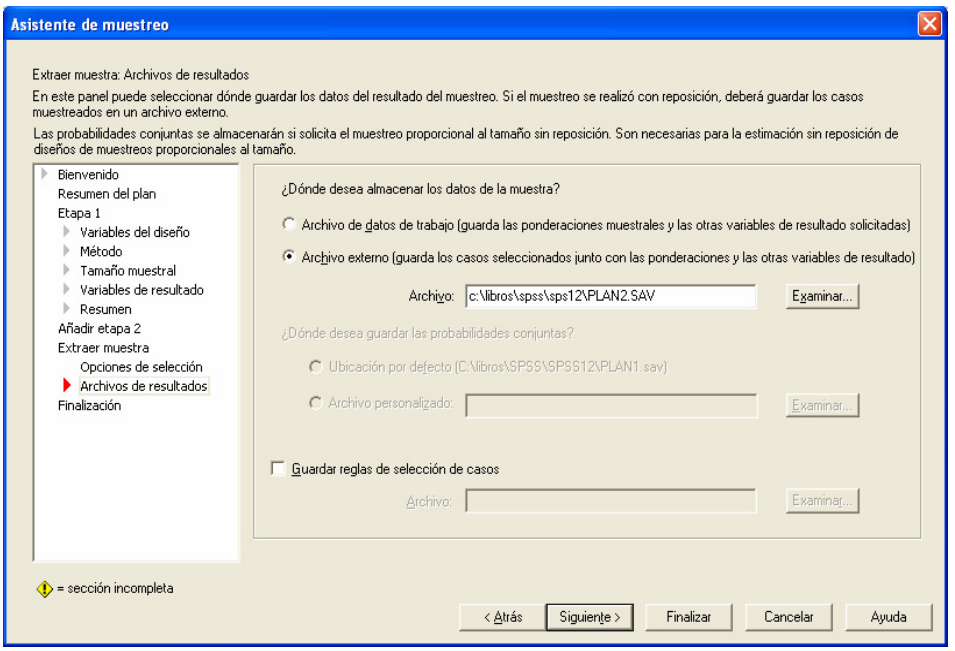


Figura 10-15

Puede ocurrir que al pulsar *Finalizar* en la Figura 10-16, algunas variables a guardar coincidan en nombre con las ya existentes. En ese caso, en la pantalla de la Figura 10-17 se hace clic en *Cambiar nombre* y SPSS realiza los cambios adecuados. La Figura 10-18 muestra el nuevo archivo PLAN2.SAV que contiene la muestra aleatoria.

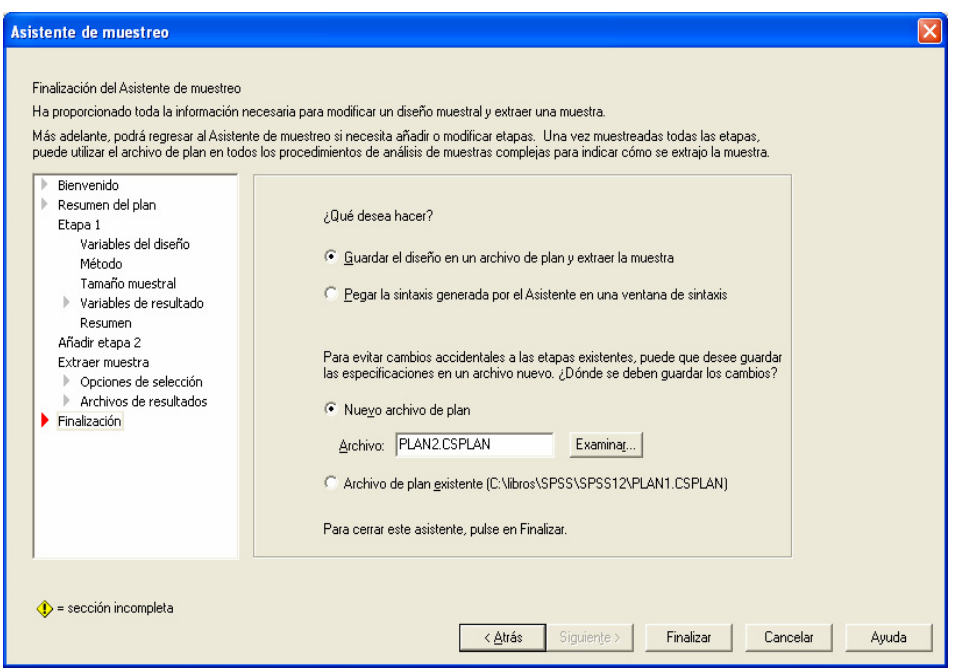


Figura 10-16

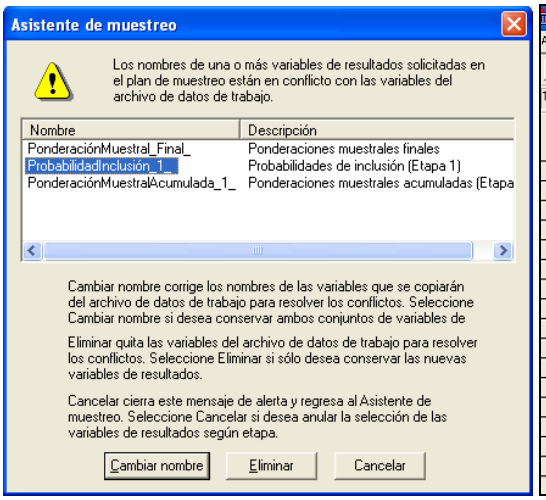
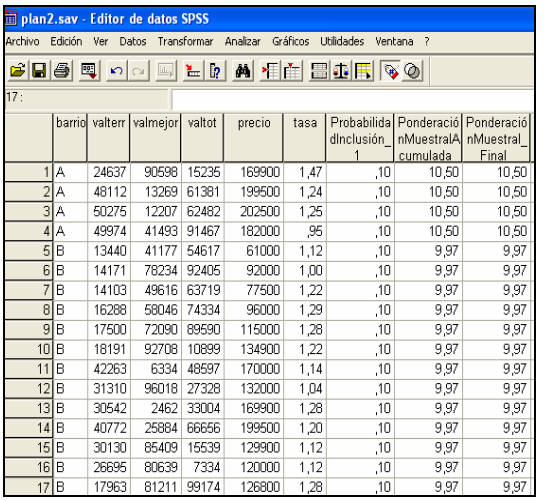


Figura 10-17



	barrio	valterr	valmejor	valtot	precio	tasa	ProbabilidadInclusión_1	PonderaciónMuestralAcumulada	PonderaciónMuestral_Final
1	A	24637	90598	15235	169900	1,47	,10	10,50	10,50
2	A	48112	13269	61381	199500	1,24	,10	10,50	10,50
3	A	50275	12207	62482	202500	1,25	,10	10,50	10,50
4	A	49974	41493	91467	182000	,95	,10	10,50	10,50
5	B	13440	41177	54617	61000	1,12	,10	9,97	9,97
6	B	14171	78234	92405	92000	1,00	,10	9,97	9,97
7	B	14103	49616	63719	77500	1,22	,10	9,97	9,97
8	B	16288	58046	74334	96000	1,29	,10	9,97	9,97
9	B	17500	72090	89590	115000	1,28	,10	9,97	9,97
10	B	18191	92708	10899	134900	1,22	,10	9,97	9,97
11	B	42263	6334	48597	170000	1,14	,10	9,97	9,97
12	B	31310	96018	27328	132000	1,04	,10	9,97	9,97
13	B	30542	2462	33004	169900	1,28	,10	9,97	9,97
14	B	40772	25884	66656	199500	1,20	,10	9,97	9,97
15	B	30130	85409	15539	129900	1,12	,10	9,97	9,97
16	B	26695	80639	7334	120000	1,12	,10	9,97	9,97
17	B	17963	81211	99174	126800	1,28	,10	9,97	9,97

Figura 10-18

ASISTENTE DE MUESTREO: EJECUTAR UN PLAN DE MUESTREO DADO

Elija en los menús *Analizar* → *Muestras complejas* → *Seleccionar una muestra...*(Figura 10-1). En el *Asistente de muestreo* seleccione *Extraer una muestra* (Figura 10-13) y elija un archivo de plan para ejecutar. Pulse *Siguiente* para continuar usando el Asistente. Revise el plan de muestreo del paso *Resumen del plan*, y a continuación pulse *Siguiente*. Cuando se ejecuta un plan de muestreo se omiten los pasos individuales que contienen información de la etapa. Ya puede pasar al paso de finalización. Si lo desea, tiene la posibilidad de especificar las etapas que ya se han muestreado.

PREPARACIÓN DE UNA MUESTRA COMPLEJA PARA SU ANÁLISIS: CREACIÓN DE UN NUEVO PLAN DE ANÁLISIS

El *Asistente de preparación del análisis* le guía a través de los pasos para crear o modificar un plan de análisis y utilizarlo con los distintos procedimientos de análisis de muestras complejas. Antes de utilizar el Asistente, debe haber extraído la muestra para el análisis de acuerdo con un diseño complejo. Es más útil crear un plan nuevo cuando no se tiene acceso al archivo del plan de muestreo utilizado para extraer la muestra (recuerde que el plan de muestreo contiene un plan de análisis por defecto). Si no tiene acceso al archivo del plan de muestreo utilizado para extraer la muestra, puede utilizar el plan de análisis contenido por defecto en el archivo del plan de muestreo u omitir las especificaciones del análisis por defecto y guardar los cambios en un archivo nuevo.

Para crear un nuevo plan de análisis, elija en los menús *Analizar muestras complejas* → *Preparar para el análisis...* (Figura 10-19), seleccione *Crear un archivo de plan* en la Figura 10-20 y elija un nombre de archivo de plan para guardar el plan del análisis. Crearemos un plan de análisis de nombre PLANA.CSAPLAN para la muestra obtenida anteriormente y guardada en el fichero PLAN2.SAV. Pulse *Siguiente* para continuar usando el Asistente. Especifique la variable que contiene las ponderaciones muestrales en el paso *Variables del diseño* y, si lo desea, puede definir estratos y conglomerados (Figura 10-21). Es posible seleccionar el método de estimación de los errores típicos en el paso *Método de estimación* (Figura 10-22). También puede especificar el número de unidades muestrales o la probabilidad de inclusión por unidad en el paso *Tamaño* (Figuras 10-23 y 10-24).

El paso *Resumen* (Figura 10-25) recoge las especificaciones de nuestro análisis. También es posible añadir una segunda o tercera etapa al diseño en el paso *Añadir etapa*. El paso *Finalización* permite guardar el archivo del plan ahora o pegar las selecciones en una ventana de sintaxis (Figura 10-26). Ahora puede pulsar *Finalizar* para guardar el plan. Se obtiene la salida del procedimiento (Figura 10-27).

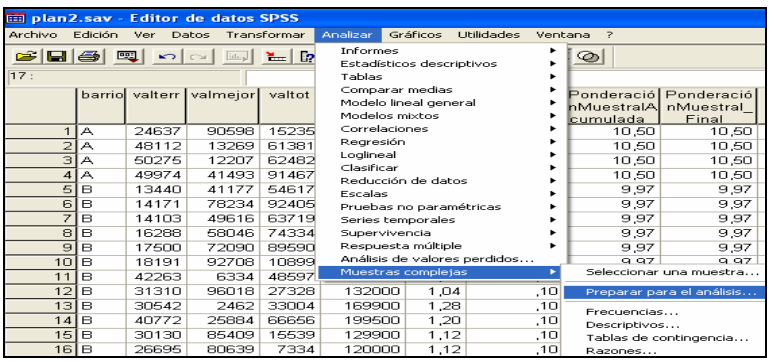


Figura 10-19

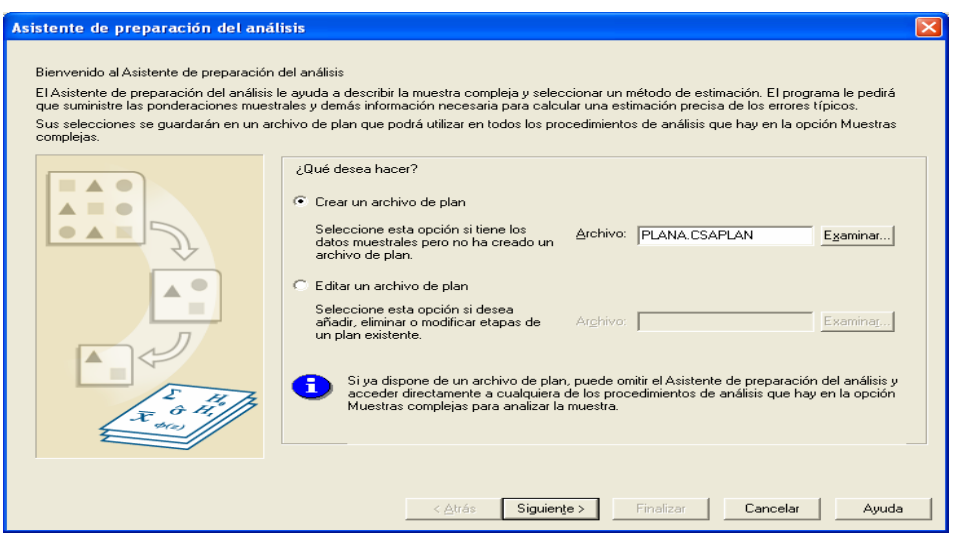


Figura 10-20

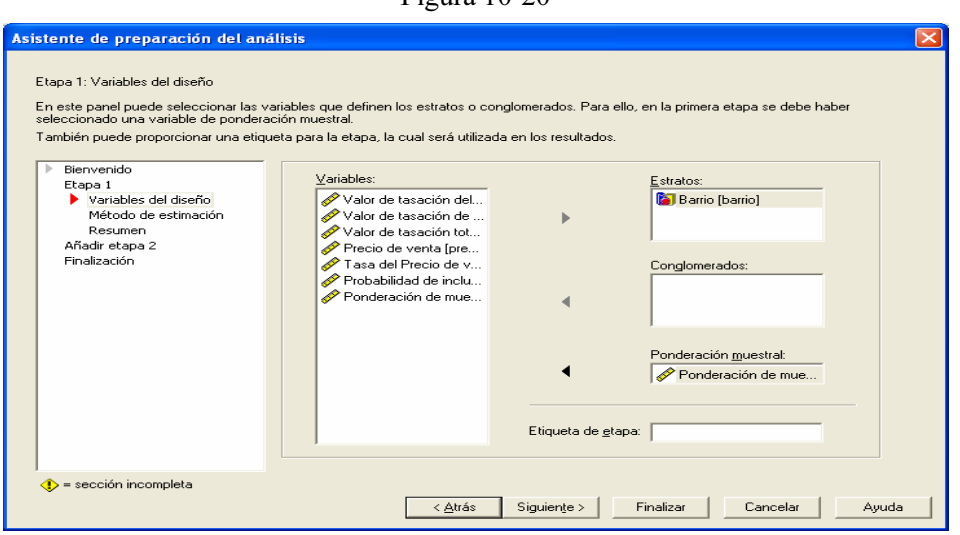


Figura 10-21

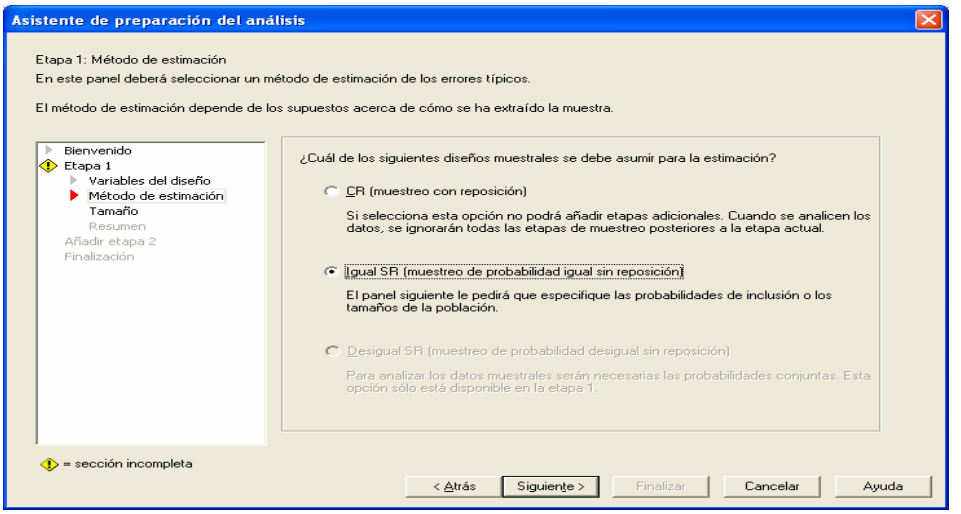


Figura 10-22

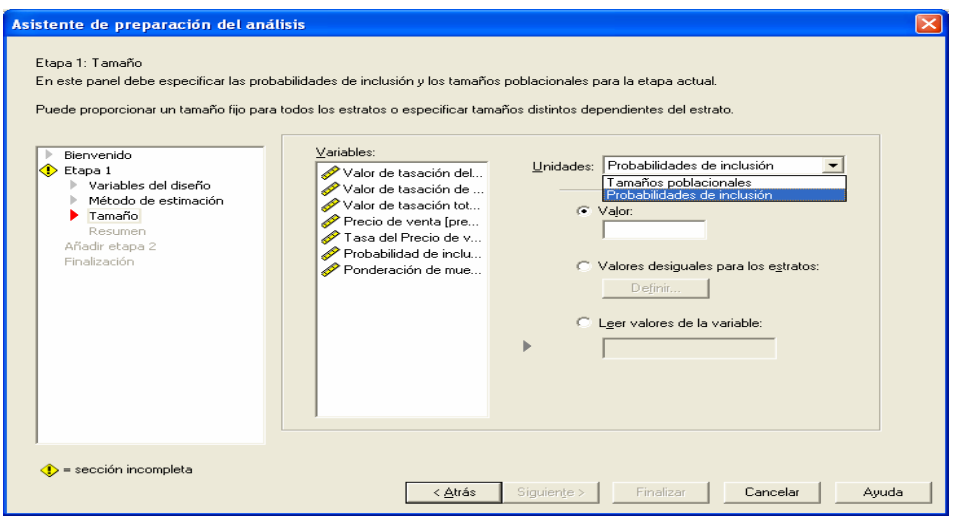


Figura 10-23

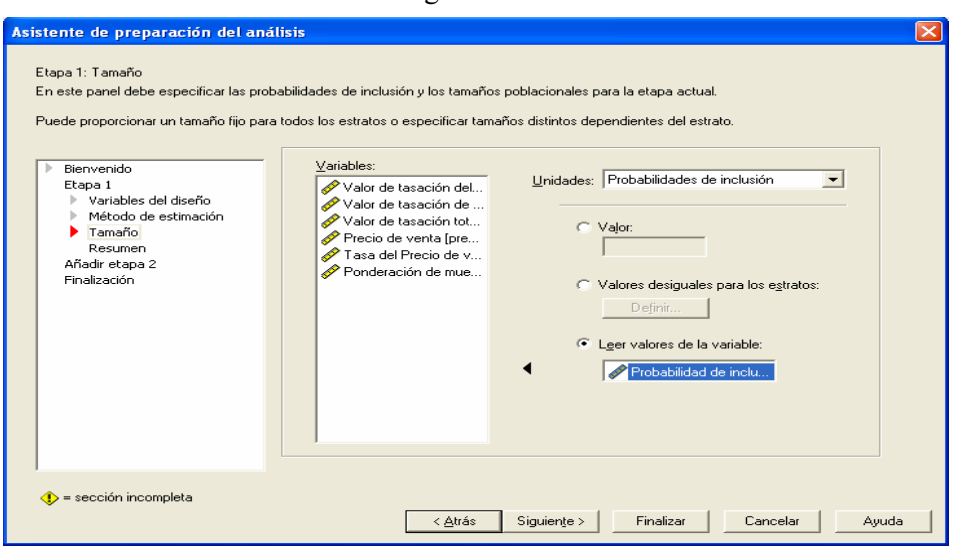


Figura 10-24

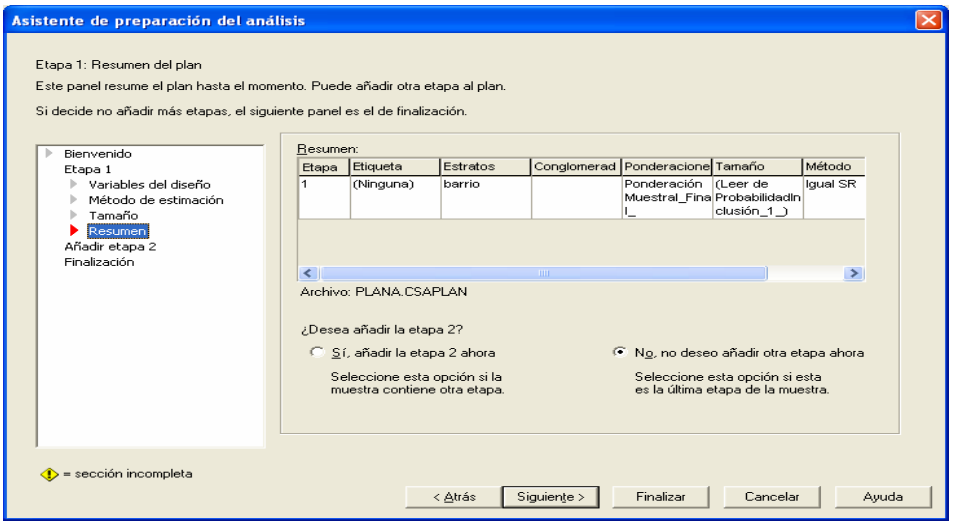


Figura 10-25

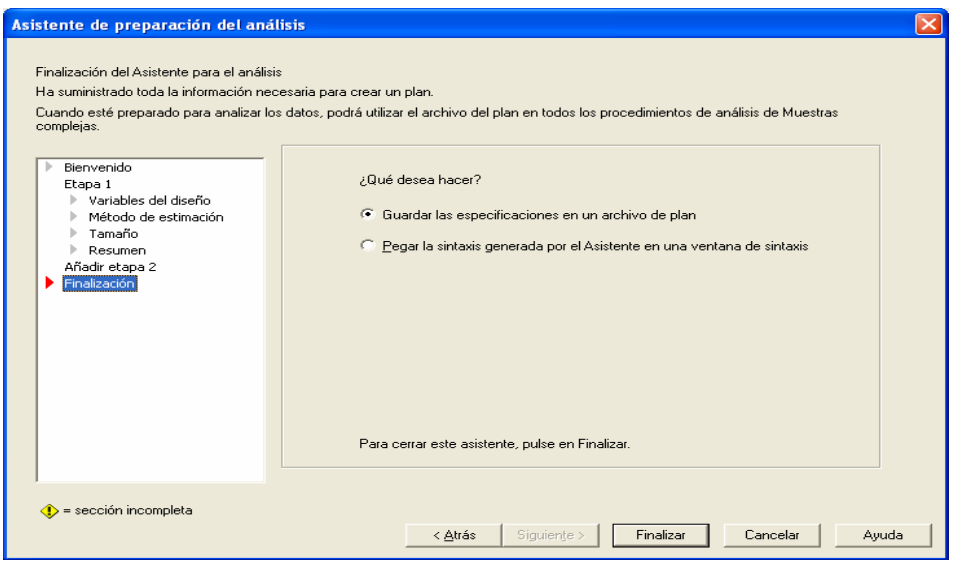


Figura 10-26

```
* Asistente de preparación del análisis.
CSPLAN ANALYSIS
 /PLAN FILE='PLANA.CSAPLAN'
 /PLANVARS ANALYSISWEIGHT=PonderaciónMuestral_Final_
 /PRINT PLAN
 /DESIGN  STRATA= barrio
   /ESTIMATOR TYPE=EQUAL_WOR
   /INCLPROB VARIABLE= ProbabilidadInclusión_1_.
```

Muestras complejas: Plan

Resumen

			Etapa 1
Variables del diseño	Estratificación	1	Barrio
Información sobre el análisis	Supuestos del estimador		Muestreo de probabilidad igual sin reposición
	Probabilidad de inclusión		A partir de la variable Probabilidad de inclusión (selección) para la etapa 1

Archivo del plan: C:\libros\SPSS\SPSS12\PLANA.CSAPLAN
Variable de ponderación: Ponderación de muestreo final

Figura 10-27

Figura 10-28

PREPARACIÓN DE UNA MUESTRA COMPLEJA PARA SU ANÁLISIS: MODIFICAR UN PLAN DE ANÁLISIS EXISTENTE

Para modificar un plan de análisis existente elija en los menús *Analizar → Muestras complejas → Preparar para el análisis*... (Figura 10-19), seleccione *Editar un archivo de plan* y elija un nombre de archivo de plan en el que se guardará el plan del análisis (Figura 10-28). Pulse *Siguiente* para continuar usando el Asistente. Revise el plan de análisis en el paso *Resumen* del plan y, a continuación, pulse *Siguiente*. Los pasos posteriores son prácticamente iguales que los de un diseño nuevo. Desplácese al paso de finalización y especifique un nombre nuevo para el archivo de plan editado o sobrescriba el archivo de plan existente. Si lo desea, tiene la posibilidad de eliminar etapas del plan.

CÁLCULOS EN MUESTRAS COMPLEJAS: FRECUENCIAS, DESCRIPTIVOS, TABLAS DE CONTINGENCIA Y RAZONES

Una vez seleccionada una muestra mediante el *Asistente de muestreo* que se activa con *Analizar→ Muestras complejas → Seleccionar una muestra...* (Figura 10-1), y preparada la muestra para su análisis mediante el *Asistente de preparación del análisis* que se activa con *Analizar→ Muestras complejas → Preparar para el análisis*... (Figura 10-19), ya estamos en disposición de calcular frecuencias, estadísticos, tablas de contingencia y razones a partir de los datos de nuestra muestra.

Frecuencias de Muestras complejas

El procedimiento *Frecuencias de Muestras complejas* genera tablas de frecuencias para las variables seleccionadas en un archivo de plan de análisis existente (*.CSAPLAN) y muestra estadísticos univariantes. Si lo desea, puede solicitar estadísticos por subgrupos definidos por una o más variables categóricas.

El procedimiento genera estimaciones de los tamaños poblacionales de las casillas, además de errores típicos, intervalos de confianza, coeficientes de variación, efectos del diseño, raíz cuadrada de los efectos del diseño, valores acumulados y recuentos no ponderados para cada estimación. Además, se calculan los estadísticos de chi-cuadrado y la razón de verosimilitudes para el contraste de proporciones de casilla iguales.

Para la obtención de *Descriptivos de Muestras complejas* elija en los menús *Analizar → Muestras complejas → Descriptivos*... (Figura 10-29), seleccione un archivo de plan, que puede ser el asociado por defecto a la muestra en memoria PLAN2.SAV u otro a especificar en *Archivo personalizado* (Figura 10-30) y, si lo desea, seleccione un archivo de probabilidades conjuntas personalizado.

Pulse en *Continuar* y seleccione al menos una variable de medida y, si lo desea, tiene la posibilidad de especificar variables para definir subpoblaciones (Figura 10-31), en cuyo caso los estadísticos se calculan por separado para cada subpoblación. Al pulsar *Aceptar* se obtiene la salida (Figuras 10-32 y 10-33).

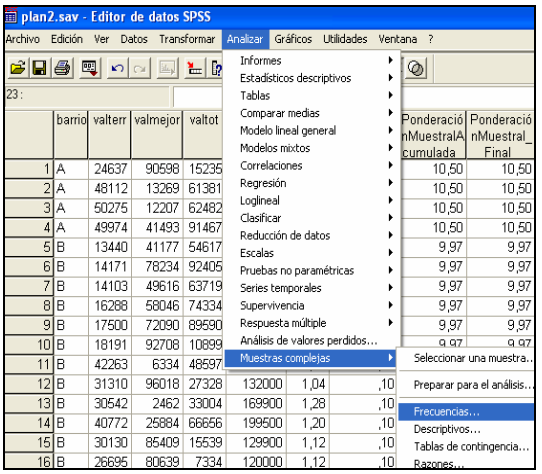


Figura 10-29

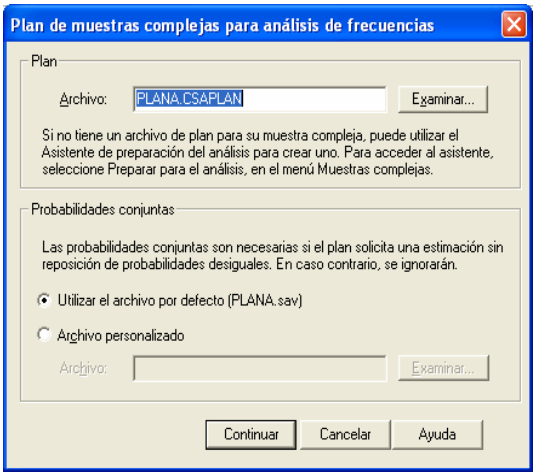


Figura 10-30

Figura 10-31

```
* Frecuencias de muestras complejas.
CSTABULATE
 /PLAN FILE = 'PLANA.CSAPLAN'
 /TABLES VARIABLES = valterr valmejor valtot precio tasa
PonderaciónMuestralAcumulada_1_
 /SUBPOP TABLE = barrio  DISPLAY=LAYERED
 /CELLS  POPSIZE
 /STATISTICS  SE
 /MISSING SCOPE = TABLE CLASSMISSING = EXCLUDE.
```

Muestras complejas: Tablas

Ponderación de muestreo acumulada para la etapa 1

		Estimación	Error típico
Tamaño de la población	9,92	258,000	,000
	9,94	467,000	,000
	9,97	319,000	,000
	10,00	500,000	,000
	10,04	482,000	,000
	10,05	372,000	,000
	10,50	42,000	,000
	Total	2440,000	,000

Figura 10-32

Tablas de subpoblación

Ponderación de muestreo acumulada para la etapa 1

Barrio			Estimación	Error típico
A	Tamaño de la población	10,50	42,000	,000
		Total	42,000	,000
B	Tamaño de la población	9,97	319,000	,000
		Total	319,000	,000
C	Tamaño de la población	9,92	258,000	,000
		Total	258,000	,000
D	Tamaño de la población	9,94	467,000	,000
		Total	467,000	,000
E	Tamaño de la población	10,00	500,000	,000
		Total	500,000	,000
F	Tamaño de la población	10,05	372,000	,000
		Total	372,000	,000
G	Tamaño de la población	10,04	482,000	,000
		Total	482,000	,000

Figura 10-33

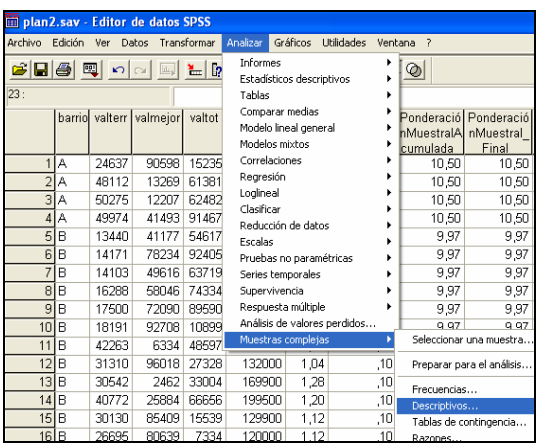


Figura 10-34

Descriptivos de Muestras complejas

El procedimiento *Descriptivos de Muestras complejas* genera estadísticos descriptivos para las variables seleccionadas en un archivo de plan de análisis existente (*.CSAPLAN). Si lo desea, puede solicitar estadísticos por subgrupos definidos por una o más variables categóricas.

El procedimiento genera estimaciones de los tamaños poblacionales de las casillas, además de errores típicos, intervalos de confianza, coeficientes de variación, efectos del diseño, raíz cuadrada de los efectos del diseño, valores acumulados y recuentos no ponderados para cada estimación. Además, se calculan los estadísticos de chi-cuadrado y la razón de verosimilitudes para el contraste de proporciones de casilla iguales. Para la obtención de *Descriptivos de Muestras complejas* elija en los menús *Analizar* → *Muestras complejas* → *Descriptivos...* (Figura 10-34), seleccione un archivo de plan, que puede ser el asociado por defecto a la muestra actual (PLANA.CSAPLAN) u otro a especificar en *Archivo personalizado* (Figura 10-35) y, si lo desea, seleccione un archivo de probabilidades conjuntas personalizado. Pulse en *Continuar* y seleccione al menos una variable de medida (Figura 10-36). Si lo desea, tiene la posibilidad de especificar variables para definir subpoblaciones, en cuyo caso los estadísticos se calculan por separado para cada subpoblación. El botón *Estadísticos* permite elegir los estadísticos a obtener (Figura 10-37). Al pulsar *Aceptar* se obtiene la salida (figuras 10-38 a 10-40).

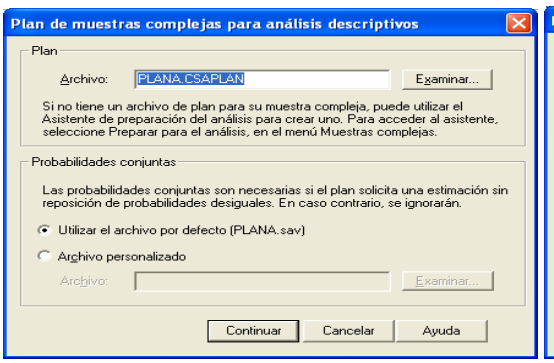


Figura 10-35

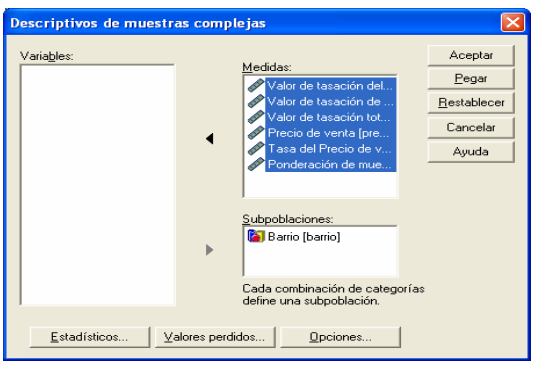


Figura 10-36

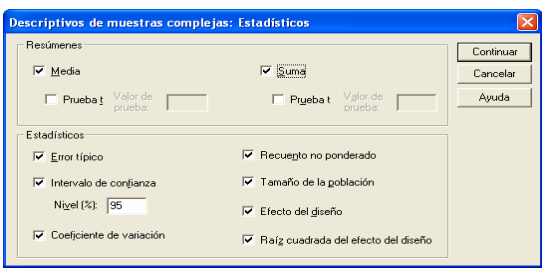


Figura 10-37

```
* Descriptivos de muestras complejas.
CSDESCRIPTIVES
 /PLAN FILE = 'PLANA.CSAPLAN'
 /SUMMARY VARIABLES =valterr valmejor valtot precio tasa
PonderaciónMuestralAcumulada_1_
 /SUBPOP TABLE = barrio  DISPLAY=LAYERED
 /MEAN
 /SUM
 /STATISTICS SE CV COUNT POPSIZE DEFF DEFFSQRT CIN (95)
 /MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.
```

Figura 10-38

Muestras complejas: Descriptivos

Estadísticos univariantes

		Estimación	Error típico	Intervalo de confianza al 95% Inferior	Intervalo de confianza al 95% Superior	Coeficiente de variación	Efecto del diseño	Raíz cuadrada del efecto del diseño	Tamaño de la población	Recuento no ponderado
Media	Valor de tasación del terreno	16441,46	501,961	15452,59	17430,34	,031	,627	,792	2440,000	244
	Valor de tasación de las mejoras	41216,69	1086,090	39077,06	43356,31	,026	,605	,778	2440,000	244
	Valor de tasación total	50293,34	1207,478	47914,58	52672,10	,024	,690	,831	2440,000	244
	Precio de venta	73553,50	2255,808	69109,50	77997,49	,031	,565	,751	2440,000	244
	Tasa del Precio de venta sobre el Valor de tasación total	1,1520	,01735	1,1178	1,1862	,015	,968	,984	2440,000	244
	Ponderación de muestreo acumulada para la etapa 1	10,0006	,00000	10,0006	10,0006	,000	,000	,000	2440,000	244
Suma	Valor de tasación del terreno	40117171	1224783,7	37704318	42530024	,031	,627	,792	2440,000	244
	Valor de tasación de las mejoras	100568720	2650060,8	95348037	1,1E+08	,026	,605	,778	2440,000	244
	Valor de tasación total	122715750	2946245,2	1,2E+08	1,3E+08	,024	,690	,831	2440,000	244
	Precio de venta	179470529	5504170,9	1,7E+08	1,9E+08	,031	,565	,751	2440,000	244
	Tasa del Precio de venta sobre el Valor de tasación total	2810,93	42,33752	2727,52	2894,33	,015	,968	,984	2440,000	244
	Ponderación de muestreo acumulada para la etapa 1	24401,57	,00000	24401,57	24401,57	,000	,000	,000	2440,000	244

Figura 10-39

Estadísticos univariantes

Barrio			Estimación	Error típico	Intervalo de confianza al 95% Inferior	Intervalo de confianza al 95% Superior	Coeficiente de variación	Efecto del diseño	Raíz cuadrada del efecto del diseño	Tamaño de la población	Recuento no ponderado
A	Media	Valor de tasación del terreno	43249,50	5918,852	31589,22	54909,78	,137	1,402	1,184	42,000	4
		Valor de tasación de las mejoras	39391,75	17469,949	4975,53	73807,97	,443	1,402	1,184	42,000	4
		Valor de tasación total	57641,25	14989,138	28112,29	87170,21	,260	1,402	1,184	42,000	4
		Precio de venta	188475,00	7291,804	174109,97	202840,03	,039	1,402	1,184	42,000	4
		Tasa del Precio de venta sobre el Valor de tasación total	1,2275	,10144	1,0277	1,4273	,083	1,402	1,184	42,000	4
		Ponderación de muestreo acumulada para la etapa 1	10,5000	,00000	10,5000	10,5000	,000	.	.	42,000	4
	Suma	Valor de tasación del terreno	1816479	248591,77	1326747	2306211	,137	,083	,289	42,000	4
		Valor de tasación de las mejoras	1654454	733737,85	208972	3099935	,443	,559	,748	42,000	4
		Valor de tasación total	2420933	629543,78	1180716	3661149	,260	,260	,510	42,000	4
		Precio de venta	7915950	306255,77	7312619	8519281	,039	,007	,084	42,000	4
		Tasa del Precio de venta sobre el Valor de tasación total	51,56	4,26042	43,16	59,95	,083	,032	,178	42,000	4
		Ponderación de muestreo acumulada para la etapa 1	441,00	,00000	441,00	441,00	,000	,000	,000	42,000	4
B	Media	Valor de tasación del terreno	23564,75	1921,755	19778,85	27350,65	,082	1,024	1,012	319,000	32
		Valor de tasación de las mejoras	56157,34	5034,367	46239,52	66075,17	,090	1,024	1,012	319,000	32
		Valor de tasación total	48472,09	4696,287	39220,30	57723,89	,097	1,024	1,012	319,000	32
		Precio de venta	120631,25	10570,260	99807,58	141454,92	,088	1,024	1,012	319,000	32
		Tasa del Precio de venta sobre el Valor de tasación total	1,1262	,02350	1,0799	1,1726	,021	1,024	1,012	319,000	32
		Ponderación de muestreo acumulada para la etapa 1	9,9688	,00000	9,9688	9,9688	,000	.	.	319,000	32
	Suma	Valor de tasación del terreno	7517155	613039,78	6309452	8724858	,082	,214	,462	319,000	32
		Valor de tasación de las mejoras	17914193	1605963,0	14750407	21077978	,090	,248	,498	319,000	32
		Valor de tasación total	15462598	1498115,6	12511274	18413922	,097	,278	,527	319,000	32
		Precio de venta	38481369	3371912,8	31838619	45124118	,088	,239	,489	319,000	32
		Tasa del Precio de venta sobre el Valor de tasación total	359,27	7,49795	344,50	374,04	,021	,017	,132	319,000	32
		Ponderación de muestreo acumulada para la etapa 1	3180,03	,00000	3180,03	3180,03	,000	,000	,000	319,000	32
C	Media	Valor de tasación del terreno	25116,88	3444,252	18331,63	31902,14	,137	1,027	1,013	258,000	26
		Valor de tasación de las mejoras	49048,65	4299,752	40578,04	57519,27	,088	1,027	1,013	258,000	26
		Valor de tasación total	62627,06	4371,214	54015,66	71236,47	,070	1,027	1,013	258,000	26
		Precio de venta	106057,69	14135,382	78210,65	133904,73	,133	1,027	1,013	258,000	26
		Tasa del Precio de venta sobre el Valor de tasación total	1,1538	,06342	1,0289	1,2788	,055	1,027	1,013	258,000	26
		Ponderación de muestreo acumulada para la etapa 1	9,9231	,00000	9,9231	9,9231	,000	,000	,000	258,000	26
	Suma	Valor de tasación del terreno	6480156	888617,05	4729559	8230753	,137	,379	,616	258,000	26
		Valor de tasación de las mejoras	12654553	1109336,1	10469134	14839971	,088	,198	,445	258,000	26
		Valor de tasación total	16157786	1127773,2	13936046	18379526	,070	,135	,368	258,000	26
		Precio de venta	27362885	3646928,6	20178348	34547422	,133	,365	,604	258,000	26
		Tasa del Precio de venta sobre el Valor de tasación total	297,69	16,36198	265,46	329,93	,055	,088	,297	258,000	26
		Ponderación de muestreo acumulada para la etapa 1	2560,15	,00000	2560,15	2560,15	,000	,000	,000	258,000	26
D	Media	Valor de tasación del terreno	18240,62	748,674	16765,71	19715,52	,041	1,010	1,005	467,000	47
		Valor de tasación de las mejoras	61016,70	1793,166	57484,12	64549,28	,029	1,010	1,005	467,000	47
		Valor de tasación total	70746,68	3250,646	64342,83	77150,53	,046	1,010	1,005	467,000	47
		Precio de venta	87459,57	2860,948	81823,44	93095,71	,033	1,010	1,005	467,000	47
		Tasa del Precio de venta sobre el Valor de tasación total	1,0781	,01766	1,0433	1,1129	,016	1,010	1,005	467,000	47
		Ponderación de muestreo acumulada para la etapa 1	9,9362	,00000	9,9362	9,9362	,000	,000	,000	467,000	47
	Suma	Valor de tasación del terreno	8518368	349630,84	7829587	9207149	,041	,097	,312	467,000	47
		Valor de tasación de las mejoras	28494800	837408,33	26845085	30144514	,029	,052	,229	467,000	47
		Valor de tasación total	33038700	1518051,5	30048102	36029298	,046	,119	,345	467,000	47
		Precio de venta	40843621	1336062,8	38211545	43475697	,033	,064	,253	467,000	47
		Tasa del Precio de venta sobre el Valor de tasación total	503,47	8,24501	487,22	519,71	,016	,017	,130	467,000	47
		Ponderación de muestreo acumulada para la etapa 1	4640,19	,00000	4640,19	4640,19	,000	,000	,000	467,000	47
E	Media	Valor de tasación del terreno	12573,40	440,769	11705,07	13441,73	,035	1,016	1,008	500,000	50
		Valor de tasación de las mejoras	39507,48	1718,870	36121,27	42893,69	,044	1,016	1,008	500,000	50
		Valor de tasación total	52080,88	2017,870	48105,63	56056,13	,039	1,016	1,008	500,000	50
		Precio de venta	59287,64	2685,951	53996,25	64579,03	,045	1,016	1,008	500,000	50
		Tasa del Precio de venta sobre el Valor de tasación total	1,1504	,03768	1,0762	1,2246	,033	1,016	1,008	500,000	50
		Ponderación de muestreo acumulada para la etapa 1	10,0000	,00000	10,0000	10,0000	,000			500,000	50
	Suma	Valor de tasación del terreno	6286700	220384,51	5852537	6720863	,035	,079	,281	500,000	50
		Valor de tasación de las mejoras	19753740	859434,84	18060633	21446847	,044	,117	,341	500,000	50

Figura 10-40

Tablas de contingencia de Muestras complejas

El procedimiento *Tablas de contingencia de Muestras complejas* genera tablas de contingencia para los pares de variables seleccionadas y muestra estadísticos sobre la clasificación bivariante. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas. Para la obtención de *Tablas de contingencia de Muestras complejas* elija en los menús *Analizar* → *Muestras complejas* → *Tablas de contingencia...* (Figura 10-41), seleccione un archivo de plan, que puede ser el asociado por defecto a la muestra actual (PLANA.CSAPLAN) u otro a especificar en *Archivo personalizado* (Figura 10-42) y, si lo desea, seleccione un archivo de probabilidades conjuntas personalizado. Pulse en *Continuar* y seleccione al menos una variable de medida para el campo *Filas* y otra para el campo *Columnas* que formarán la tabla de contingencia (Figura 10-43). Si lo desea, tiene la posibilidad de especificar variables para definir subpoblaciones, en cuyo caso las tablas se calculan por separado para cada subpoblación. El botón *Estadísticos* permite elegir los estadísticos a obtener (Figura 10-44). Al pulsar *Aceptar* se obtiene la salida.

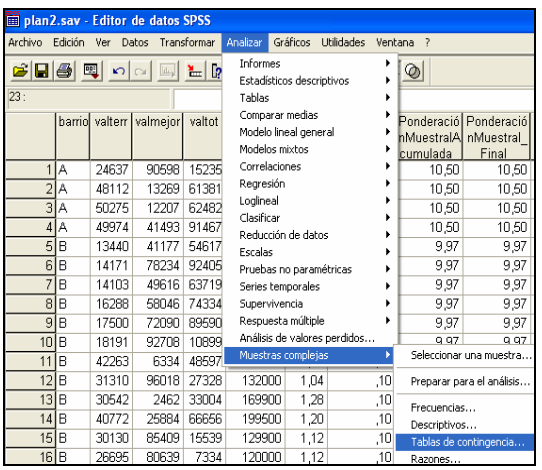


Figura 10-41

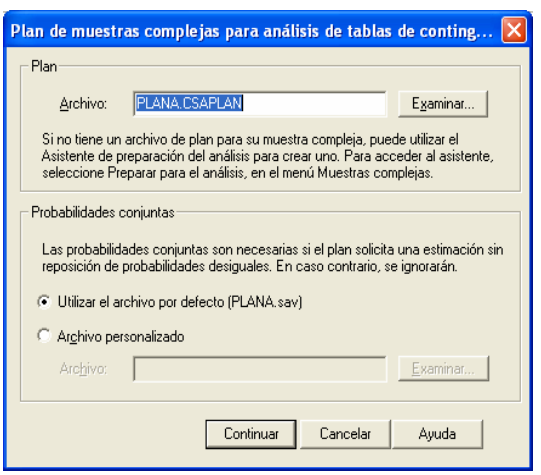


Figura 10-42

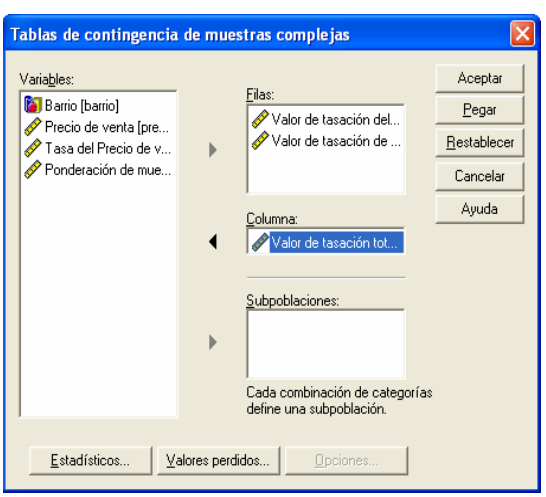


Figura 10-43

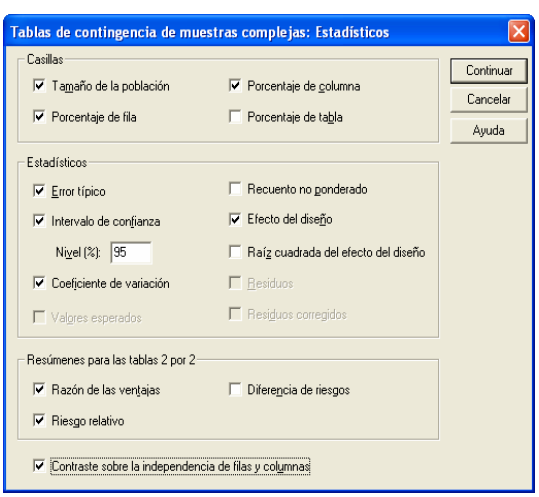


Figura 10-44

Razones de Muestras complejas

El procedimiento *Razones de Muestras complejas* muestra estadísticos de resumen univariantes para razones de variables. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas. Para la obtención de *Razones de Muestras complejas* elija en los menús *Analizar* → *Muestras complejas* → *Razones...* (Figura 10-45), seleccione un archivo de plan, que puede ser el asociado por defecto a la muestra actual (PLANA.CSAPLAN) u otro a especificar en *Archivo personalizado* (Figura 10-46) y, si lo desea, seleccione un archivo de probabilidades conjuntas personalizado. Pulse en *Continuar* y seleccione al menos una variable de medida para el campo *Numerador* y otra para el campo *Denominador* que formarán la razón a estimar (Figura 10-47). Los numeradores y los denominadores deben ser variables de escala con valores positivos. Si lo desea, tiene la posibilidad de especificar variables para definir subpoblaciones, en cuyo caso, las razones se calculan por separado para cada subpoblación.

El botón *Estadísticos* permite elegir los estadísticos a obtener (Figura 10-48). Al pulsar *Aceptar* se obtiene la salida (Figuras 10-49 y 10-50).

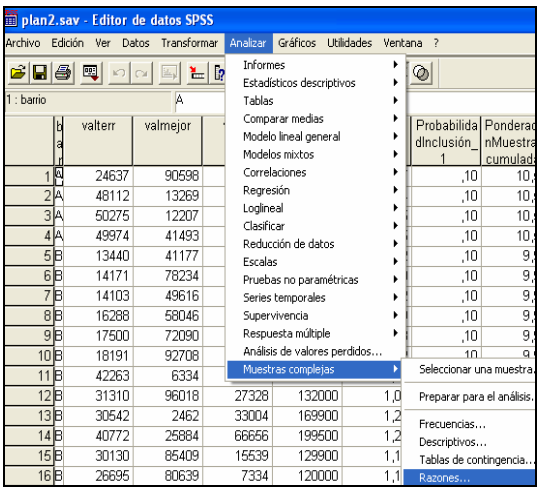


Figura 10-45

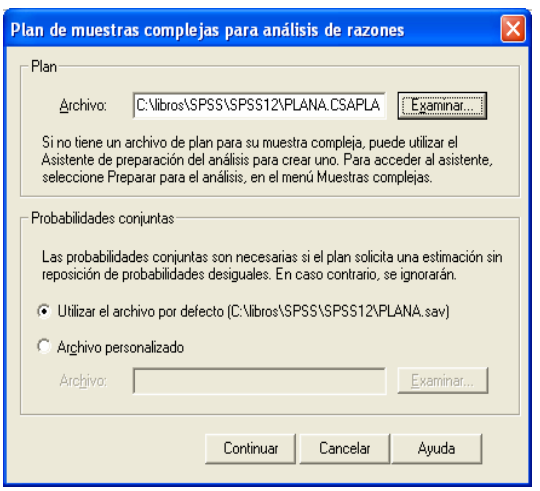


Figura 10-46

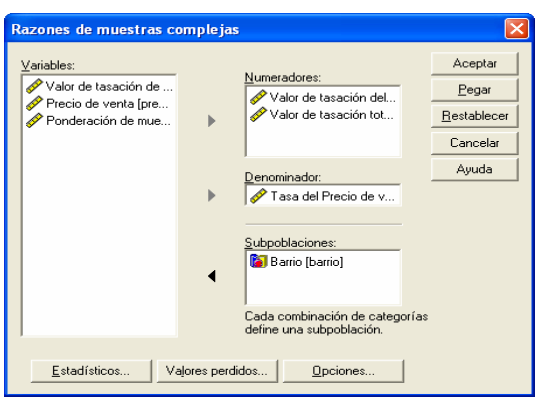


Figura 10-47

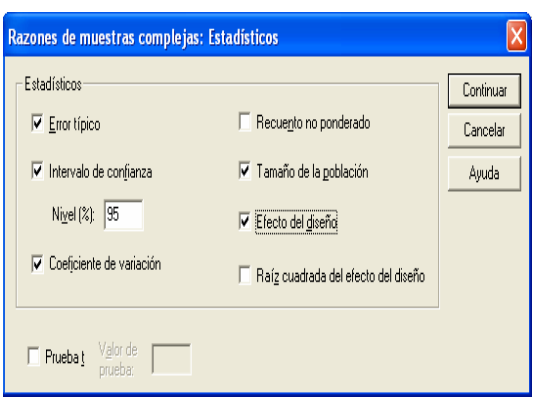


Figura 10-48

```
* Razones de muestras complejas.
CSDESCRIPTIVES
 /PLAN FILE = 'C:\libros\SPSS\SPSS12\PLANA.CSAPLAN'
 /RATIO NUMERATOR = valterr valtot  DENOMINATOR = tasa
 /STATISTICS SE CV POPSIZE DEFF CIN (95)
 /SUBPOP TABLE = barrio  DISPLAY=LAYERED
 /MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.
```

Muestras complejas: Descriptivos

Razones 1

Numerador	Denominador	Estimación de la razón	Error típico	Intervalo de confianza al 95%		Coeficiente de variación	Efecto del diseño	Tamaño de la población
				Inferior	Superior			
Valor de tasación del terreno	Tasa del Precio de venta sobre el Valor de tasación total	14271,865	482,914	13320,513	15223,218	,034	,669	2440,000
Valor de tasación total	Tasa del Precio de venta sobre el Valor de tasación total	43656,684	1281,198	41132,694	46180,675	,029	,734	2440,000

Figura 10-49

Razones 1

Barrio	Numerador	Denominador	Estimación de la razón	Error típico	Intervalo de confianza al 95%		Coeficiente de variación	Efecto del diseño	Tamaño de la población
					Inferior	Superior			
A	Valor de tasación del terreno	Tasa del Precio de venta sobre el Valor de tasación total	35233,809	7311,077	20830,811	49636,806	,208	1,402	42,000
	Valor de tasación total	Tasa del Precio de venta sobre el Valor de tasación total	46958,248	15996,197	15445,355	78471,141	,341	1,402	42,000
B	Valor de tasación del terreno	Tasa del Precio de venta sobre el Valor de tasación total	20923,196	1667,242	17638,690	24207,703	,080	1,024	319,000
	Valor de tasación total	Tasa del Precio de venta sobre el Valor de tasación total	43038,485	4149,755	34863,369	51213,601	,096	1,024	319,000
C	Valor de tasación del terreno	Tasa del Precio de venta sobre el Valor de tasación total	21767,967	3085,760	15688,945	27846,988	,142	1,027	258,000
	Valor de tasación total	Tasa del Precio de venta sobre el Valor de tasación total	54276,800	5357,548	43722,302	64831,298	,099	1,027	258,000
D	Valor de tasación del terreno	Tasa del Precio de venta sobre el Valor de tasación total	16919,459	854,113	15236,835	18602,083	,050	1,010	467,000
	Valor de tasación total	Tasa del Precio de venta sobre el Valor de tasación total	65622,538	2893,095	59923,071	71322,005	,044	1,010	467,000
E	Valor de tasación del terreno	Tasa del Precio de venta sobre el Valor de tasación total	10929,590	507,370	9930,059	11929,121	,046	1,016	500,000
	Valor de tasación total	Tasa del Precio de venta sobre el Valor de tasación total	45271,975	2462,091	40421,596	50122,354	,054	1,016	500,000
F	Valor de tasación del terreno	Tasa del Precio de venta sobre el Valor de tasación total	10840,512	658,576	9543,102	12137,923	,061	1,030	372,000
	Valor de tasación total	Tasa del Precio de venta sobre el Valor de tasación total	30226,807	1944,133	26396,818	34056,795	,064	1,030	372,000
G	Valor de tasación del terreno	Tasa del Precio de venta sobre el Valor de tasación total	7884,094	532,573	6834,912	8933,276	,068	1,022	482,000
	Valor de tasación total	Tasa del Precio de venta sobre el Valor de tasación total	27733,296	2402,642	23000,034	32466,558	,087	1,022	482,000

Figura 10-50